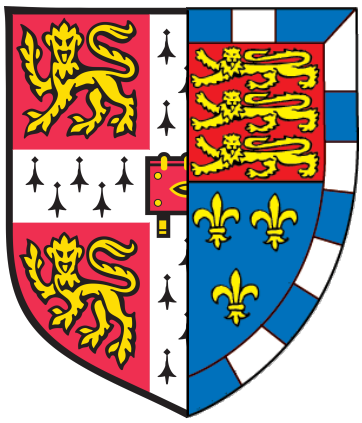


Comparative and Epidemiological Genomics of Human Malaria Parasites

Gavin George Rutledge
of
Christ's College, Cambridge
and
the Wellcome Sanger Institute

This dissertation is submitted
for the degree of
Doctor of Philosophy

University of Cambridge
Cambridge, United Kingdom
September 2018

1. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

2. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

3. It does not exceed the prescribed word limit for the relevant Degree Committee.

Internal Supervisors:
Dr. Matthew Berriman
Prof. Dominic P. Kwiatkowski

External Supervisors:
Dr. Thomas D. Otto
Dr. Colin A. Russell

Comparative and Epidemiological Genomics of Human Malaria Parasites

by Gavin G. Rutledge

Following both significant advances and setbacks in the past decades of fighting malaria, the end goal of malaria elimination is now once again within sight. However, this endgame may prove the most challenging yet, as there are still significant gaps in our understanding of human malaria more generally and as we know from past experience that the malaria parasite is able to rapidly overcome any challenge thrown at it. Despite the huge international endeavour to understand the genomic basis of malaria biology, the genome sequences of two of the five human malaria parasite species, *Plasmodium malariae* and *P. ovale*, have remained essentially a mystery. Consequently, the implications of these sequences on aspects such as drug resistance have eluded us. However, even for human malaria parasite species that have been sequenced at large scale, such as *P. falciparum*, a better understanding of the impact of genetic variation on drug resistance is needed, especially in light of multidrug resistance in Southeast Asia. In this thesis, I have, in collaboration with others, explored these different aspects of human malaria parasites, showing to what extent sequencing data can inform our understanding of human malaria and aid us in our fight against this devastating disease.

Initially I assembled reference genome sequences for both *P. malariae* and *P. ovale*, an analysis of which I present in Chapter 1. I show that the *P. malariae* genome is markedly different to other *Plasmodium* genomes and relate this to its unique biology. Using additional draft genome assemblies, I also confirm that *P. ovale* consists of two species that appear to have diverged millions of years ago.

Internal Supervisors:
Dr. Matthew Berriman
Prof. Dominic P. Kwiatkowski

External Supervisors:
Dr. Thomas D. Otto
Dr. Colin A. Russell

In Chapter 2, I use the newly assembled *P. malariae* reference genome in combination with clinical data to characterize a case of clinical recrudescence of a *P. malariae* infection, and suggest that drug resistance may have played a role. To better understand the ability of malaria parasites to acquire drug resistance, in Chapter 3 I harnessed a large dataset of *P. falciparum* whole genome sequences with associated phenotype data on mefloquine, an antimalarial drug. I show that the current outbreak of multidrug resistance in Southeast Asia is accompanied by a hyper-sensitization of the parasite population through the acquisition of a complex genetic architecture of mefloquine sensitivity. Finally, in Chapter 4, by incorporating phenotype data on additional drugs, including chloroquine, artemisinin and piperaquine, I identify a specific haplotype of the *pfcrt* gene that displays super-resistance to chloroquine and that acts as a genetic backbone to artemisinin resistance and to multidrug resistance in general.

The approach taken in this thesis is one of extracting new information from layering on additional data. I begin by comparing genome sequences to each other in Chapter 1, I then layer on clinical metadata in Chapter 2, I add in phenotype data for one drug in Chapter 3, and finally, in Chapter 4, I harness phenotype data for multiple drugs. At each level, I identify new biology that both expands our understanding of human malaria in general and sheds light on the specifics of antimalarial drug resistance. The contributions made in this work will be of significant importance in the upcoming end game of malaria elimination.

Contents

Dedicated to my family.

Acknowledgments

First and foremost, I wish to thank all my PhD supervisors. Thank you to Matt Berriman for giving me the freedom and independence to explore data and to make progress at my own pace. Whenever I came up with a seemingly crazy analysis or needed more sequencing data, I knew that I could count on you to say 'go right ahead'. Thank you to Dominic Kwiatkowski for showing me how large-scale global science can be performed and how to see the forest for the trees. You have taught me how to see the big picture of the science that we do. Thank you to Colin Russell for putting my over-ambitious aims into perspective and challenging me on defining my concrete goals. Finally, thank you to Thomas Otto for his continued support throughout my PhD, from pitching the initial plan of the PhD project to me, to helping with the technical analyses and paper writing, as well as providing emotional support and friendship. Your enthusiasm and passion for science were infectious and brought the best out of me.

I also wish to thank Roberto Amato for his guidance on navigating the huge datasets we have on hand and for helping me define the narrative of the analyses I had performed. I wish to thank Chris Newbold for all the invigorating discussions we had every Wednesday on all things malaria, they were one of many highlights in my PhD.

Thank you to everyone in team 112 and 133. I am glad I was able to spend time in both groups during my PhD and thereby got to know such a large group of interesting people.

I also wish to thank the Wellcome Trust and the Medical Research Council for their generous funding that has enabled me to pursue this PhD. Thank you to the graduate programme at the Sanger In-

stitute for their support from the beginning to the end of the PhD. I wish to thank Christ's College for the warm community that was like a second home away from home.

I wish to thank my parents for their continued support, love, and understanding. It is the values and principles that you have instilled in me that have enabled me to come this far.

Finally, thank you to my wife, Ruijiao Liu, for always standing by my side, in good times as in bad times. A PhD is like a roller-coaster ride, with moments of pure joy when analyses bear fruit and bouts of soul-crushing agony when these are subsequently scooped by competitors. It has been you who has had to endure all my mood swings, who has always consoled me, and who has encouraged me to push on and to do the best I can do. Thank you.

A PhD is a difficult undertaking, fraught with pitfalls and setbacks. I have therefore been very lucky to have had amazing people to advise and support me throughout. Thank you everyone.

...and in summer dysenteries, diarrhoeas, and protracted quartan fevers frequently seize them, and these diseases when prolonged dispose such constitutions to dropsies, and thus prove fatal.

Hippocrates, On Airs, Waters, and Places, 400 BCE

0

An Introduction to Malaria

0.1 The Basics of Malaria

0.1.1 Global Burden of Malaria

Malaria is a life-threatening disease caused by intra-erythrocytic parasites of the *Plasmodium* genus. According to the World Health Organisation (WHO), there were an estimated 216 million clinical cases of malaria in 2016, resulting in almost half a million deaths[368]. While malaria is a disease that is found globally (figure 1), with indigenous cases having been reported in 91 countries across multiple

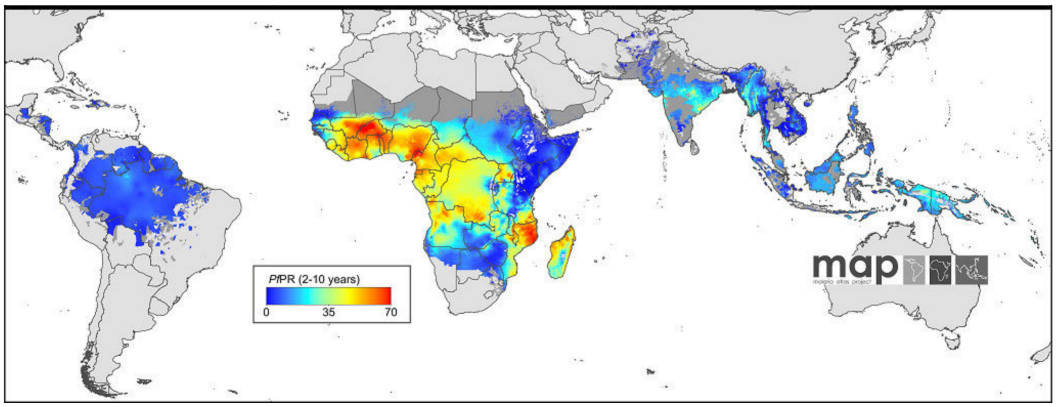


Figure 1: Global distribution of P. falciparum malaria in 2010. Figure adapted from Gething et al. [118].

continents in 2016 [16], the vast majority of these malaria cases (> 90%) occurred in Africa [368]. This distribution also coincides with the number of malaria fatalities, with 91% of malaria fatalities in 2016 being estimated to have occurred on the African continent [368].

At the outset of this thesis, there were five recognized species of *Plasmodium* parasites that infect humans, including *P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale*, and the zoonotic *P. knowlesi*. However, there had been suggestions that *P. ovale* may consist of two distinct species, *P. o. wallikeri* and *P. o. curtisi* [339], which would imply the existence of six human-infective malaria parasite species. The different species differ significantly in terms of their public health impact because of differing levels of prevalence and disease severity. Subsequently, research efforts and resources have been unequally distributed across these species, with some human-infective species, such as *P. malariae* and *P. ovale*, having been largely neglected by the research community.

P. falciparum is by far the most well-studied human-infective species. It causes the majority of malaria-related deaths and is consequently the prime subject of most malaria research. *P. falciparum* is found across all malaria-endemic regions (figure 1) and there were an estimated 207 million cases of *P. falciparum*-associated malaria in 2016 [368]. Most of these *P. falciparum* malaria cases (>90%) occurred

in sub-Saharan Africa, where malaria transmission remains very high[16]. The case fatality rate of *P. falciparum* malaria is estimated at about 0.256%, which is about 7-times higher than the estimated fatality rate of the second most important human-infective species, *P. vivax* (0.0375% case fatality rate)[368]. This high fatality rate is due to the unique propensity of *P. falciparum* to cause infected red blood cells to clump together with each other and even with uninfected red blood cells. These clumps can then 'sequester' in the host's microvasculature causing obstruction[16]. The most severe and fatal cases of malaria are the result of this obstruction happening in the brain, known as cerebral malaria[361].

While *P. vivax* malaria is less fatal than *P. falciparum* malaria, it is the most important form of malaria outside of Africa. There were an estimated 8.5 million clinical cases of *P. vivax* malaria globally in 2016[368]. *P. vivax* is almost absent from most regions of Africa as the parasite requires the Duffy antigen on the red blood cell surface for successful invasion, while the human population in Africa mostly lacks the Duffy antigen[220]. Some recent studies however suggest that *P. vivax* may in rare cases successfully invade Duffy negative individuals[217]. Outside Africa, *P. vivax* cases exceed those of *P. falciparum*, especially in South America[16]. *P. vivax* is less deadly than *P. falciparum*[368], but is capable of forming a dormant liver stage, called a hypnozoite[231]. Dormant hypnozoites can result in malaria relapses many months after the initial infection. This aspect of *P. vivax* malaria significantly increases the morbidity associated with it[11].

The most recently discovered human-infective species is *P. knowlesi*, a species that is usually restricted to infecting macaques. While macaques are the natural host of *P. knowlesi*, it can be transmitted to humans as a zoonosis[16]. *P. knowlesi* has quite a restricted geographical range, as the macaques it infects are limited to Malaysia. Initially, *P. knowlesi* infections were misdiagnosed as *P. malariae* infections due to their morphological similarity[325]. *P. knowlesi* infections are frequently fatal if un-

treated, as the parasite can propagate quickly within the host leading to high levels of parasitaemia and consequently severe anaemia[72]. The case fatality rate of *P. knowlesi* is thought to be similar to that of *P. falciparum* if the infection is not recognized quickly[289]. While the overall burden of *P. knowlesi* infections is not well understood, it seems that infections with the species may be more common in the Malaysian region than initially thought[72,24].

The first *Plasmodium* species to be viewed under a microscope was *P. malariae* and was hence named after the disease. It is widespread and is found throughout most malaria-endemic regions[68]. While present in all these regions, *P. malariae*-associated malaria is relatively benign and is therefore likely to be under-reported[16]. Many reported cases of *P. malariae* result from patients presenting to the hospital with *P. falciparum* or *P. vivax* malaria and then discovering to be co-infected with *P. malariae*[298]. *P. malariae* infections are however not completely benign, as they may occasionally result in fatal renal complications[178]. Furthermore, *P. malariae*, even though it is not thought to produce hypnozoites, has the ability to cause recrudescences many years after the initial infection[318]. This ability to remain hidden in the host for decades is not well understood, but can result in *P. malariae* infections being a lifelong disability if not treated.

The final human-infective species is *P. ovale*, which may possibly consist of two morphologically indistinguishable subspecies[339]. *P. ovale* is found throughout Africa and Asia, but it is conspicuously absent in South America[16]. It is not known why *P. ovale* is not found in South America. Similar to *P. malariae*, infections with *P. ovale* are relatively benign and the species is frequently only reported in co-infections with *P. falciparum* and *P. vivax*[298]. Similar to *P. vivax*, *P. ovale* forms dormant hypnozoites and can result in malaria relapses years after initial infection[67]. The species is not well understood and has only recently been proposed to consist of two separate species, *P. o. curtisi* and *P. o.*

wallikeri[339,250,10]. Both of these have been shown to co-occur in the same region[249] and even within the same host[110]. There have been suggestions that *P. o. curtisi* may have longer relapse times than *P. o. wallikeri*[244], but little else is known to distinguish the two biologically or clinically.

0.1.2 General Malaria Biology

Plasmodium parasites are eukaryotic organisms belonging to the phylum of protozoan parasites known as the *Apicomplexa*, thereby being related to other infectious agents such as *Toxoplasma gondii*, causing toxoplasmosis, *Babesia* species, the agents of babesiosis, *Cryptosporidium parvum*, the parasite causing cryptosporidiosis, and many others[292]. The *Apicomplexa* are single-cellular organisms characterized by the presence of an invasion-related apical complex, consisting of numerous subcomponents such as rhoptries and micronemes[189]. Another distinguishing feature of most species of *Apicomplexa*, including the *Plasmodium* species, is the presence of a unique organelle termed the apicoplast[214], thought to have originated from an ancestral event of secondary endosymbiosis of a red algae resembling *Chromera*[196]. All *Apicomplexa* have complex lifecycles, involving several stages of multiplication and changes between haploid and diploid states[189], however the *Plasmodium* lifecycle is particularly complex due to the obligate need for both a host and a vector to complete its full lifecycle.

Human-infective *Plasmodium* species rely on both the human host and on a mosquito vector to complete their lifecycle (figure 2). Species of *Plasmodium* parasites vary in the specific mosquito species that are able to transmit them[33]. The most common human malaria vector species are mosquitoes belonging to the *Anopheles gambiae* species complex, however over 50 species have been identified as competent human malaria vectors[136]. When an infected mosquito bites a human to take a blood meal, hundreds of sporozoites, a haploid and motile stage of the malaria parasite, are injected through the

mosquito saliva into the human skin[285] (figure 2). Upon entering the blood stream, the sporozoites migrate to the human liver, where they infect hepatocytes and are virtually invisible to the immune system[345]. For two days to three weeks following liver infection, a process of rapid asexual multiplication termed merogony occurs within the infected hepatocytes[79]. Each of these infected hepatocytes, now termed a schizont, results in the generation of hundreds-to-thousands of infective merozoites[285]. In most human-infective malaria parasite species, upon completion of merogony all the infected hepatocytes release the merozoites into the bloodstream by budding off vesicles known as merosomes[335] (figure 2). Two species of human-infective malaria parasite species, *P. vivax* and *P. ovale*, may result in the formation of a dormant liver stage, termed a hypnozoite[175]. Hypnozoites can persist for years within the host and often result in relapses long after the initial infection has been cleared[317,210]. The process by which hypnozoites are reactivated is not fully understood[210,312], however the result is that the merozoites within the hypnozoite are released into the bloodstream and thereby restart a blood-stage infection.

The merozoites that are released into the bloodstream now attempt to invade host erythrocytes (figure 2). The process by which the merozoites identify and ultimately invade the fast-flowing target erythrocytes, all the while surviving in a highly hostile environment, is rapid, completed in under a minute[363]. Merozoites initially attach to erythrocytes using merozoite surface proteins (MSPs), such as MSP-1[71] (figure 3). The parasites then makes use of two classes of adhesins, Duffy binding proteins (DBPs)[137] and reticulocyte binding proteins (RBPs, known as Rh proteins in *P. falciparum*)[113]. The adhesins bind to specific host proteins, such as glycophorins[48,207,212] in the case of DBPs, and their differential distribution across species has been suggested to contribute in part to the cell specificity of the different species[113,225], with *P. falciparum*, *P. knowlesi* and *P. malariae* invading normocytes (ma-

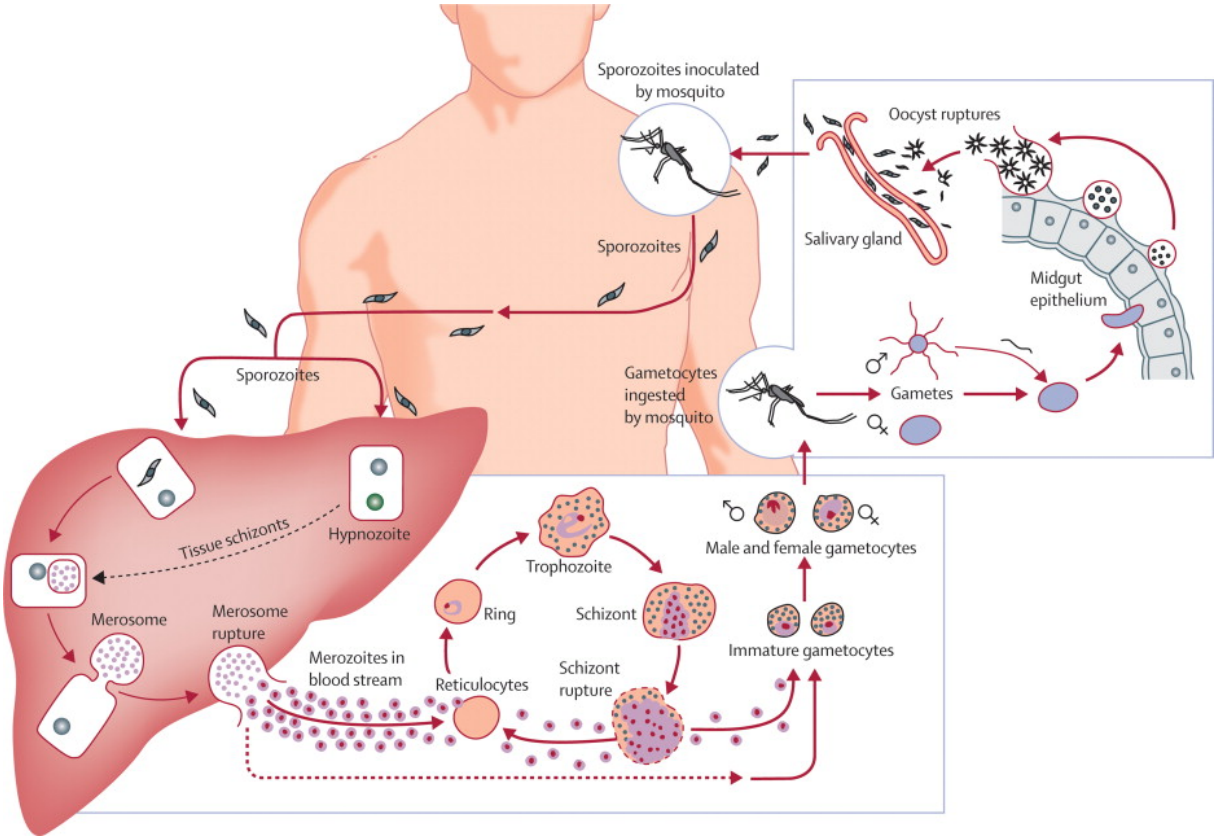


Figure 2: Plasmodium lifecycle overview. Figure reproduced from Mueller et al. [231].

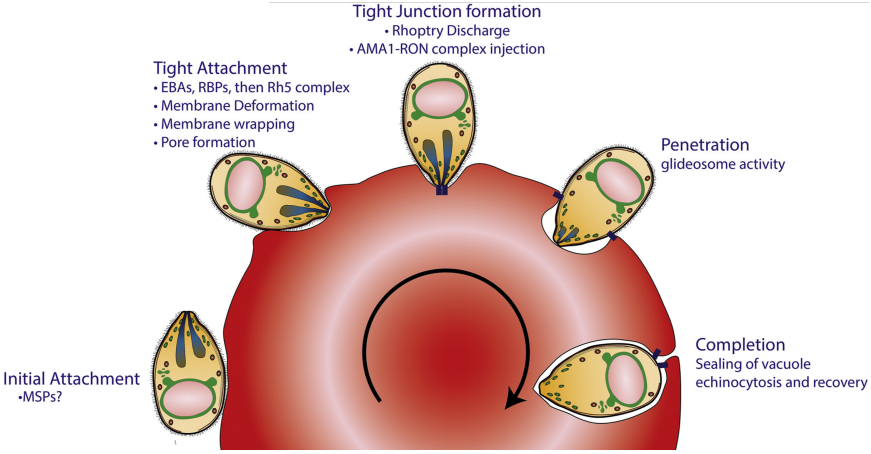


Figure 3: Overview of erythrocyte-invasion steps. Figure reproduced from Cowman et al. [71].

ture erythrocytes) and with *P. vivax* and *P. ovale* preferring reticulocytes (young nucleated erythrocytes)[155,225]. Most adhesins are thought to be dispensable individually, however the overall function they perform is essential to invasion[202]. In particular, one RBP in *P. falciparum*, PfRh5, is essential to invasion through its binding to human basigin[74]. The PfRh5 protein is thought to be tethered as a tripartite complex of PfRh5 with PfRIPR and CyRPA to the GPI-anchored Pf113[360,111], where engagement of the complex with basigin would initiate pore formation between the parasite and erythocyte membranes[363]. This essential role played by PfRh5 has therefore made it a prime vaccine target[273]. Upon pore formation, a tight junction is formed between apical membrane antigen 1 (AMA1) on the parasite surface and the RON complex that the parasite inserts into the erythrocyte membrane[347] (figure 3). The parasite then utilizes this anchoring to propel itself into the erythrocyte, losing its surface proteins on the way, and finally sealing the vacuole behind, producing the parasitophorous vacuole that the parasite resides in intra-erythrocytically[71].

Upon completion of invasion, the parasite begins another process of asexual replication within the red blood cell, known as schizogony. This particular asexual replication cycle lasts approximately 48

hours in most species, however *P. malariae* takes 72 hours and *P. knowlesi* takes only 24 hours to complete it. The cycle is often split into three phases that are morphologically and metabolically distinct: the ring stage, the trophozoite stage, and the schizont stage (figure 2). The ring-stage, named so by its appearance when Giemsa-stained, is the initial stage following invasion. The parasite begins feeding on the erythrocyte's haemoglobin, catabolising it into a haem derivative (ferriprotoporphyrin IX) that is then converted into inert haemozoin crystals[23]. As the parasite grows within the red blood cell, it eventually morphs into a trophozoite with a more rounded shape, an increased number of ribosomes, and an enlarged rough endoplasmic reticulum, allowing for increased protein synthesis[23]. As the parasite ramps up its protein production, it eventually undergoes multiple nuclear division, schizogony, producing between 6 and 32 merozoites that take up most of the intra-erythrocytic space[91]. The number of merozoites that a schizont harbours varies by species, with *P. malariae* and *P. ovale* generally having fewer than *P. falciparum*, which in turn produces fewer merozoites per schizont than *P. vivax* (Centre for Disease Control, US). Finally, these newly formed merozoites egress from the red blood cell through a protease degradation of the parasitophorous vacuole and erythrocyte membrane[35]. As they egress, the merozoites search for new red blood cells to infect, thereby restarting the intra-erythrocytic cycle[120].

Throughout the intra-erythrocytic cycle, the parasite exports a number of proteins to the red blood cell surface in order to avoid the host immune system, modifying the morphology of the erythrocyte in the process[78]. The malaria parasite has two main methods in which it alters the red blood cell to avoid detection by the immune system[122]. Firstly, malaria parasite species express highly variable antigenic proteins that may enable long-lasting infections by preventing a unilateral immune response against any specific surface protein[98,306]. Using these antigenic proteins to direct the immune re-

sponse against a subset of infected red blood cells is thought to keep the parasite population from killing the host, thereby enabling the establishment of a chronic infection [102]. Secondly, certain proteins that are transported to the erythrocyte membrane are involved in either binding to other red blood cells [247], resulting in a process called rosetting where multiple infected and uninfected red blood cells clump together [140], or in binding to the vascular endothelium, enabling the infected red blood cell to sequester in the microvasculature, thereby avoiding clearance by the spleen [270].

It is this intra-erythrocytic cycle that results in the clinical disease symptoms of malaria. The usually synchronised egress of merozoites from the infected red blood cells results in a strong immune response that presents itself as a high fever (usually referred to as a paroxysm), leading to the characteristic tertian fever (every three days for most human malaria species) or quartan fever (every four days for *P. malariae*) that is often used to initially diagnose the disease [114]. *P. knowlesi* produces a quotidian fever due to its shorter lifecycle. The continuous feeding on red blood cells can lead to severe anemia in chronic infections [130], as well as splenomegaly from the spleen removing all the infected red blood cells [45]. Severe malaria cases involve malaria parasites sequestering in the brain, referred to as cerebral malaria, which is fatal in almost 20% of cases [232]. Similarly, malaria infections during pregnancy can result in parasites sequestering in the placenta, resulting in miscarriages, low birth weights, and maternal anemia [226]. The latter two complications are restricted to *P. falciparum* infections and, while most malaria deaths are due to severe anemia [232], they do contribute in part to the high mortality associated with this species [16].

While most infected red blood cells will develop into schizonts that release merozoites to continue the blood infection, a certain proportion of infected erythrocytes develop into the sexual stage of the malaria parasite, a process termed gametocytogenesis [159] (figure 2). Infected red blood cells that

undergo gametocytogenesis develop into either male or female gametocytes, known as micro- and macrogametocytes respectively. The rate of commitment, *ie.* the proportion of infected red blood cells that develop into sexual stages, is not fully understood[80,2], but certain environmental stressors such as anemia or drug treatment are known to contribute to higher levels of commitment[53]. While most human-infective malaria parasite species complete gametocytogenesis within 2 days and produce round gametocytes, *P. falciparum* takes a lengthy 10-12 days to complete the process[159] and produces sickle shaped gametocytes (from which the species derives its name).

As a mosquito feeds on an infected host, it may ingest by chance both a male and a female gametocyte as part of its blood meal (figure 2). The change in environment from human blood to mosquito mid-gut, *ie.* change in temperature, pH and exposure to xanthurenic acid, causes the gametocytes to mature to gametes[34]. These haploid gametes then fuse into a diploid ookinete with meiotic recombination occurring in the process[324]. Ookinetes traverse the epithelial mid-gut wall to form oocysts[159]. Within each oocyst, thousands of sporozoites develop that eventually egress and travel to the mosquito's salivary glands via the haemocoel[296]. The sporozoites remain in the salivary glands until the mosquito takes a new blood meal, at which point the sporozoites are injected into the host to restart a malaria infection[324].

0.1.3 Global Malaria Eradication Program

Malaria has afflicted humans for thousands of years, and while humans fought back against the disease on a genetic level by developing resistance mutations[164], such as that resulting in sickle cell anemia[4], it was in the early 20^{th} century that humankind put up a real fight to eliminate the disease. Malaria was long known to be common around marshes, with 'bad air' (latin: *mal-aria*) being thought of

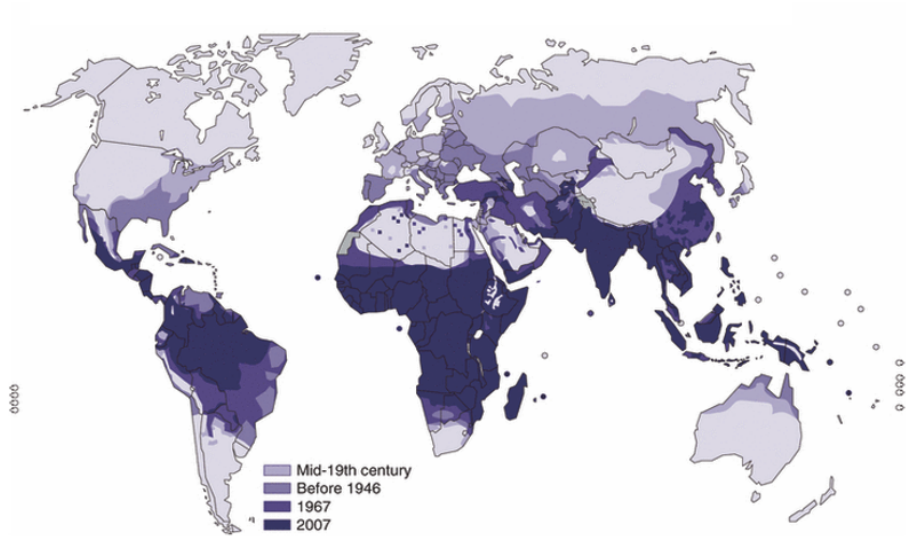


Figure 4: Changes in global malaria endemicity from the mid-19th century to the 21st century. The change in endemicity from 'Before 1946' to '1967' can largely be attributed to the GMEP. Figure reproduced from Kamini et al. [162].

as the cause of the disease. As the mechanisms of the disease, such as transmission via mosquitoes, were finally discovered at the turn of the 20th century, control methods including draining marshes, deploying bed nets and the use of the antimalarial drug quinine led to the elimination of malaria from a number of regions, mostly parts of North America and Europe [162] (figure 4). Heartened by successful elimination efforts in a number of countries, together with the advent of the first residual insecticide, dichloro-diphenyl-trichloroethane (DDT), and the synthesis of the highly effective antimalarial drug, chloroquine, the WHO was given the mandate in 1955 to direct a global campaign to eradicate malaria: the Global Malaria Eradication Program (GMEP)[237].

The GMEP was a highly ambitious project that aimed for the complete eradication of malaria within a decade. Based on positive experiences of malaria elimination in a limited number of epidemiological settings, the GMEP was based on the concept of complete coverage of in-door residual spraying (IRS) of DDT to disrupt transmission [101]. This complete coverage of DDT was advocated at

the expense of other traditional control interventions, such as marsh draining and bed nets[237]. Significant inroads were initially made, for instance by eradicating malaria from Europe and North America, as well as by significantly reducing the incidence in certain countries such as India (figure 4). In 1963 however, funding for the program began to diminish as the US stopped contributing, resulting in available resources becoming stretched. The progress of the GMEP began to stall, and, around 1967, certain countries reverted from the post-elimination 'consolidation phase' back to the on-going transmission 'attack phase'. For example, Sri Lanka (then known as Ceylon) had virtually eliminated malaria by 1963 leading them to halt their IRS of DDT, which in turn led to a strong resurgence of malaria in the following years, culminating in large epidemics in 1968 and 1969[165]. These local setbacks were exacerbated by the emergence of mosquitoes with resistance to DDT in a number of regions, as well as the discovery of the negative environmental impact of DDT spraying. In addition to this, the single-minded focus on reducing transmission by using DDT meant that the appearance of *P. falciparum* parasites that exhibited chloroquine resistance in pockets of Southeast Asia and in South America in the 1950's was not given the attention it deserved[237]. By the end of the 1960's, chloroquine resistance had spread throughout those regions, eventually reaching East Africa in the 1970's and the rest of Africa by the mid 1980's[272]. In light of all of these setbacks, and despite the initial successes of the GMEP, the program was abandoned in 1969 as it became apparent that eradication was not possible any more.

0.1.4 Current Progress

The abandonment of the GMEP in 1969 is often said to have succeeded in eradicating, instead of malaria, the job prospects of any aspiring malariologists at the time, as malaria research was largely

neglected for many years after the campaign was called off[237]. The WHO recommended a switch to control efforts in order to maintain the gains that were made during the GMEP, resulting in resources being unevenly distributed with a disproportionate amount being designated for regions with very low levels of malaria. With this approach, almost no gains were made in further reducing the global limits of malaria until a shift in policy with the launch of the Roll Back Malaria (RBM) movement[135]. The RBM campaign, launched in 1998, aimed to halve malaria mortality, morbidity, and economic burden of the disease by 2010[235]. As opposed to the control efforts that preceded it, the RBM movement focused resources on areas that were highly endemic for malaria[135]. The movement was able to make significant inroads[328], and over time additional funding partners joined the effort, such as the Bill and Melinda Gates foundation. The RBM partnership has now published their current goal of eliminating malaria by 2030 as part of the Action and Investment to defeat Malaria 2016-2030 (AIM) plan (www.rollbackmalaria.com). This plan of eliminating malaria by 2030 is reminiscent of the ambitious aims of the GMEP, and it is on that backdrop that current elimination efforts are taking place.

Over the last two decades, the RBM movement has resulted in almost halving the number of malaria-associated deaths[62,123], and the number of countries that are endemic to malaria have also continued to decrease over the years (figures 5 & 6)[368]. This reduction in malaria deaths and prevalence can largely be attributed to the use of insecticide-treated bednets (ITN), in-door residual spraying (IRS), and the use of artemisinin-based antimalarial drugs[32]. However, the WHO has recently stated in its 2017 World Malaria Report that 'Progress appears to have stalled.' and that '...in some countries and regions, we are beginning to see reversals in the gains achieved.'[368]. While this has to do with both a stalling in increasing ITN coverage and a reduction in IRS[368], insecticide and antimalarial drug resistance also play a role and threaten the progress made to date. Pyrethroid-resistance in the

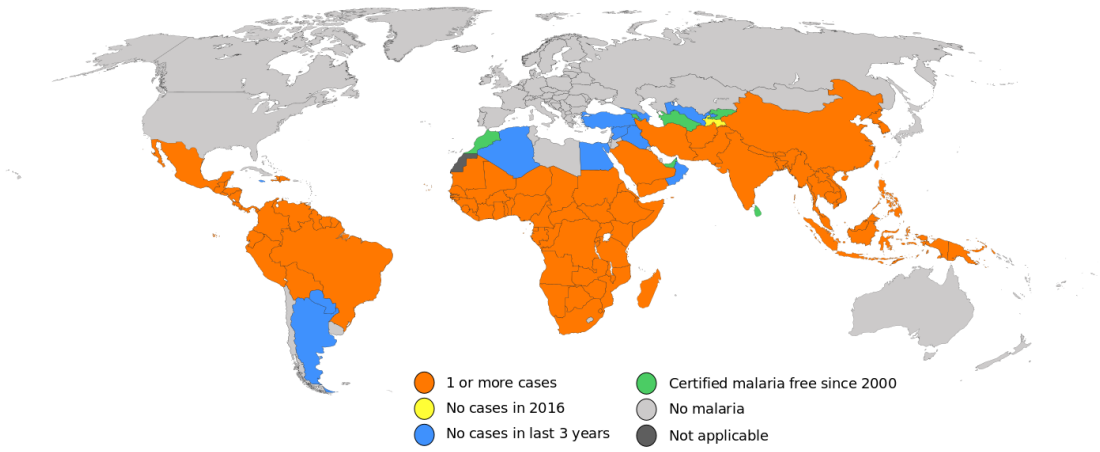


Figure 5: Overview of changes in malaria endemicity by country in 2016. Figure reproduced from the WHO World Malaria Report 2017 [368].

mosquito vector is now widespread and at high frequency throughout most of the malaria-endemic regions of the African continent, which is coincidentely also where ITNs and IRS are the primary public health interventions [138]. While effective alternatives such as piperonyl butoxide (PBO) treated bednets exist [284], they are very slow to be deployed at scale [167].

While insecticide-resistance is already harming progress and potentially leading to increased cases of malaria, the threat of antimalarial drug resistance also looms on the horizon. The front-line antimalarial drug treatment for the last two decades has been artemisinin combination therapies (ACTs), deployed across all malaria-endemic regions for the treatment of *P. falciparum* malaria [368], while other drugs such as chloroquine are occasionally still used for non-falciparum malaria [378]. Due to the short-acting half-life of artemisinin, it is usually administered with a long-acting partner drug such as piperaquine or mefloquine, with the intent of slowing down the acquisition and spread of resistance to the drugs [134]. This is based on the idea that the parasites that are able to survive the short-acting

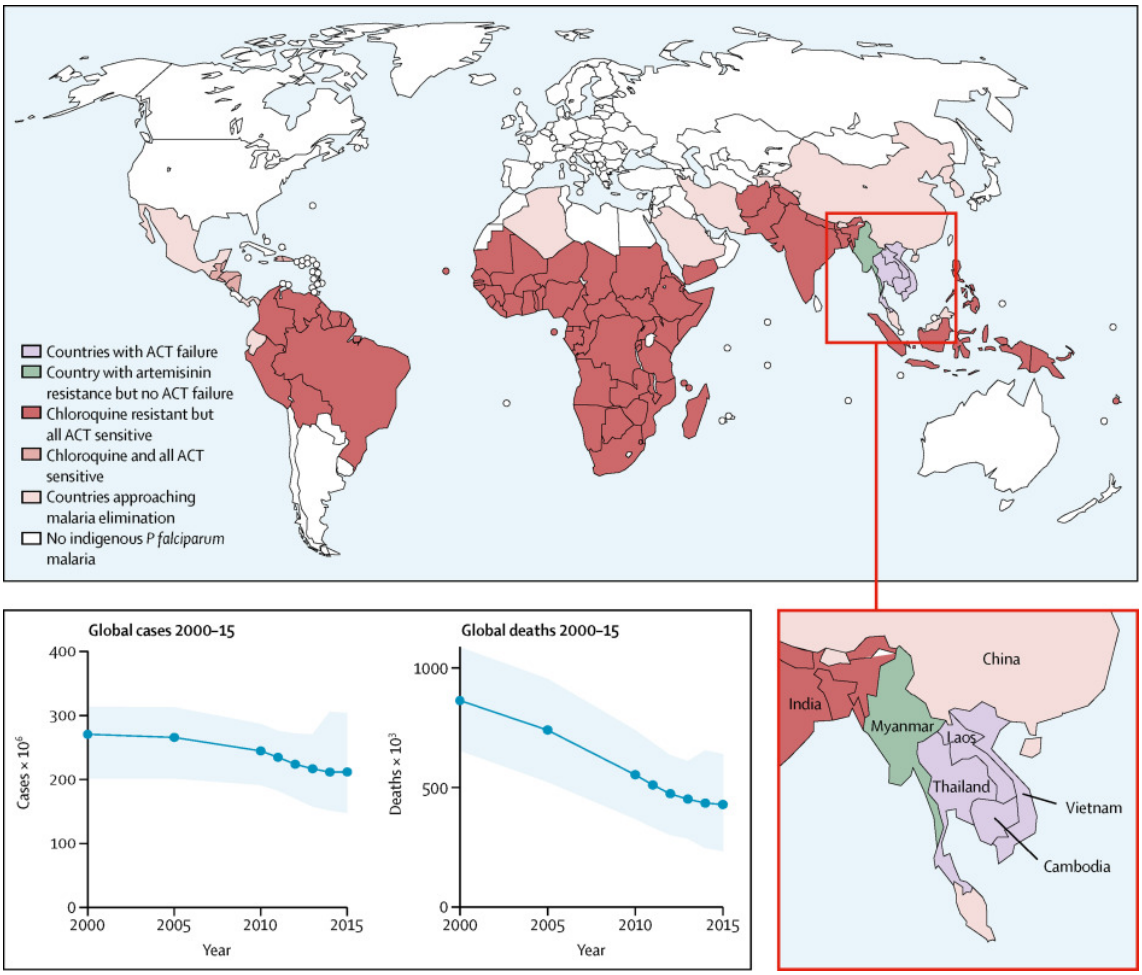


Figure 6: Global distribution of antimalarial drug resistance in P. falciparum (top and bottom-right), as well as change in the number of cases and deaths attributed to P. falciparum from 2010 to 2015 (bottom-left) . Figure reproduced from Ashley et al. [16].

artemisinin, either by being resistant or by being in a lifecycle stage that is unaffected by the drug, would subsequently be cleared by the longer-acting partner drug[302]. Despite these precautions, parasite resistance to artemisinin was first reported in 2008 in Western Cambodia[243] and has subsequently been observed throughout the Southeast Asian region[84,15,351] (figure 6). The slow-clearing artemisinin-resistant parasites were initially still cleared by the partner drug, which is piperaquine in these particular countries, however in recent years these resistant parasites appear to also have acquired resistance to piperaquine, leading to cases of complete treatment failure[185,187]. The threat that now overshadows current control efforts is the possibility of these multidrug resistant parasites spreading to Africa, where most malaria cases occur but no resistance to artemisinin has yet to be reported[374].

The current situation is in many ways analogous to that of chloroquine resistance during the time of the GMEP, where tragically chloroquine resistance ended up spreading from Southeast Asia to Africa, leading to innumerous deaths. We are, as the WHO says[368], 'at a crossroads'. With complacency and a reduction in funding, a reenactment of the failed GMEP may be on the cards, however with continued effort by all stakeholders involved, we may learn from history and avoid losing the ground that we have gained. Many things are different now to the way they were in the 1960's, as our understanding of malaria has increased tremendously. We also have one very special new tool in our toolkit, the topic of which this thesis is about, namely genomics.

0.2 Malaria Genomics Introduction

0.2.1 The Road to Sequencing Malaria Parasites

Sequencing technologies have advanced at a rapid pace over the last two decades and there have been a number of breakthroughs that have transformed our ability to characterize malaria parasites on a genomic level that have led up to this. An important development early on in enabling genetic studies of *Plasmodium* parasites was the adaptation of *P. falciparum* to continuous *in vitro* culture[348], opening up an endless supply of parasites to experiment on. The ability of *P. falciparum* to infect normocytes enabled this culturing and is what has prevented, for instance, *P. vivax* from being cultured, as it is restricted to reticulocytes, which are difficult to obtain in the lab[313]. Recently, *P. knowlesi* has been successfully adapted to continuous *in vitro* culture[224] and is now often used as a proxy for *P. vivax* as they are more closely related to each other than to *P. falciparum*[97]. Attempts to culture either *P. ovale* and *P. malariae* have until now been unsuccessful[67,68].

For species that cannot be cultured *in vitro*, such as *P. vivax*, as well as to study clinical samples or understand local population structure, advances have been made in collecting and preparing samples for sequencing directly from the field. The main limitation had previously been the low amount of parasite DNA that can be extracted from a sample compared to the high abundance of host (*ie.* human) DNA. Indeed, within a whole blood sample from a patient with a 1% parasitaemia, there will be roughly 10 times more parasites than white blood cells (WBC) but due to the human genome being 300 times larger than the parasite genome, only approximately 5% of the DNA material will be of parasite origin[21]. To circumvent this problem, protocols have been developed to deplete WBC[21] from samples or to enzymatically degrade the host DNA[262]. Another obstacle has been the low yield

of DNA from field samples, especially for the samples with low parasitaemia levels. PCR-based whole genome amplification (WGA) protocols, some especially designed for the low-GC biased *P. falciparum* genome, are now commonly used to amplify the amount of DNA in those samples[264,263]. Furthermore, both problems of low DNA yield and high abundance of host DNA have recently been tackled using a unified method called selective whole genome amplification (sWGA), which makes use of specially designed primers that bind to sequence motifs that are significantly enriched in the parasite genome compared to the host genome[188]. These different advancements now make it possible to perform whole genome sequencing of malaria parasites directly from dried blood spots (DBS)[261].

The advances in preparing the parasite DNA for sequencing have been complemented by the rapid advances in sequencing technologies over the last two decades. The first type of sequencing technology to be widely used for *Plasmodium* sequencing was capillary-based 'Sanger' sequencing, an expensive technology that produced long sequencing reads of up to 1kb (kilobase) in length and had a very low rate of sequencing errors[305]. Sanger sequencing was used for many of the original reference genome sequencing projects, including among others the human genome[153], the *P. falciparum* genome[115], and the *Anopheles gambiae* genome[144]. While the technology has fallen out of favour in recent years due to the high costs associated with it, it is still sometimes used for confirmation of highly polymorphic regions due to its low error rate.

In the mid 2000's, massively-parallel 'next generation' sequencing (NGS) technologies were developed that had shorter reads but significantly higher throughput at a much lower cost than Sanger sequencing, with innovators in the area including Solexa (acquired by Illumina)[28], 454 (acquired by Roche)[88], and SOLiD sequencing by Life technologies[316]. The fierce competition in the area drove innovation and sequencing costs dropped by four orders of magnitude between 2007 and 2012[314]. Il-

lumina now dominates the market for short-read sequencing with a range of different sequencing machines. These short-read sequencers produce sequencing reads of up to 300bp (base pairs) in length. The short reads have proven very useful in aligning them back to reference genome sequences to study genetic variation (see below), however using them to perform *de novo* genome assemblies has proven challenging [315].

The original Sanger sequencing reads enabled genome sequences to be assembled via an 'overlap-layout-consensus' method, using greedy assembly softwares such as the TIGR assembler [340] or the Celera CAP3 [146], a method which required an all-against-all pairwise similarity check comparison between all the input reads. These methods therefore scaled exponentially with the number of input reads, causing data storage and run-time memory requirement problems with the huge volume of short reads coming from the NGS sequencers. This meant that new approaches had to be developed to handle this huge volume of data, a problem for which there is no efficient solution [233]. Graph-based assembly tools formalize the problem by treating reads as strings and then attempting to find the minimal superstring that contains all the strings using the concept of a string-graph [234]. Many assemblers utilizing a string graph prune the graph for spurious connections to speed up the assembly process, but still tend to scale suboptimally with the volume of read input. An innovation was the development of a method similar to a string-graph, but based on *k-mers* (read fragments of size *k*) that are connected in a so-called *de Bruijn* graph. This *k-mer* based approach is able to scale well with the volume of input reads, as identical *k-mers* are collapsed into a single node in the *de Bruijn* graph [278]. Assemblers employing a *k-mer*-based approach include among others Velvet [379], ABySS [323] and SOAP-denovo [206]. Hybrid assemblers utilizing a combination of these approaches also exist, such as the MaSuRCA assembler [383]. By fragmenting the reads, *k-mer* based approaches have difficulties

handling repetitive regions. These difficult regions can frequently be overcome using paired end (PE) sequencing, where two reads originate from the same original sequencing template with a certain gap (known as the fragment size) between them. Standalone software, such as SSPACE[37], is able to use these paired reads to 'scaffold' assembly gaps caused by repetitive regions, though most assemblers also have in-built scaffolding processes.

In addition to the short-read sequencers by Illumina, in recent years a number of long-read sequencing machines have been developed by companies such as Pacific Biosciences (PacBio)[95] and Oxford Nanopore Technologies (ONT)[183]. The former is based on optical reading of polymerase-mediated synthesis in real time using fluorescently labelled nucleotides[95], while the latter utilizes the flow of ions that a DNA molecule emits when it moves through a nanometre-scale pore in order to infer the nucleotide sequence[183]. Hence, both technologies do not require template amplification, leading to the generation of large sequencing reads often exceeding 10kb in length, with some reads from the ONT almost reaching 1Mb (megabase)[156]. While the throughput for both technologies is still below that of the high-throughput Illumina sequencers, they are making rapid progress[314]. Both technologies are also known to have high rates of sequencing errors (>10%), which are fortunately randomly distributed in PacBio (and can therefore be corrected for with sufficient coverage), but may be for now more systematic for ONT[314]. On the other hand, the ONT sequencer, known as the minION, has the added advantage of being extremely portable, being the size of a USB stick, which has made it an extremely valuable tool for sequencing samples in the field[141]. The long reads generated by these sequencing technologies are uniquely useful for performing *de novo* genome assemblies as they are able to span large homopolymer tracks and other repetitive regions[279]. With the reduction in throughput and increased read length, specialised long-read assemblers are often based on string-graphs, including

HGAP[59] or HINGE[160] for example, though *k-mer*-based approaches also exist, such as CANU[172]. Finally, hybrid assemblers that can use short reads and long reads together have also recently been developed, with examples including HySA[100] and the updated MaSuRCA assembler[384].

0.2.2 *P. falciparum* Genome Sequencing

In 1996, a consortium of genome centres and funders was established to sequence the complete genome sequence of *P. falciparum*[142]. This multi-year project eventually completed in 2002 with a slew of publications[117,41,132,116,148] describing the 23Mb genome sequence of *P. falciparum* clone 3D7[115]. The extremely low GC content of the genome sequence (<20% GC) caused both experimental and computational difficulties, such as DNA fragments not being stable in *Escherichia coli* and large AT repeats causing difficulties in the assembly process[142]. While the first version of the *P. falciparum* reference genome was published in 2002, there has been continuous maintenance and curation of the genome sequence and annotation by the Wellcome Sanger Institute in the UK, with the current version, 3.1 (released in August 2015), having added the 35kb apicoplast genome sequence for instance. The 3D7 reference genome sequence together with the curated annotation is publicly available and searchable through GeneDB (www.genedb.org) and PlasmoDB (www.plasmodb.org).

The ≈23.3Mb *P. falciparum* 3D7 reference genome sequence consists of 15 contigs, making up the 14 chromosomes and the apicoplast genome sequence, but missing the corresponding mitochondrial genome sequence[115]. The current assembly is of very high quality, with no sequencing gaps and with the chromosome sequences extending from telomere to telomere. The chromosomes range in size from chromosome 1 (≈0.64Mb) to chromosome 14 (≈3.29Mb)[115]. The overall AT content of the genome is extremely high with 82% AT, though this is unevenly distributed with higher levels in the

'core' region of the genome and lower levels in the 'subtelomeric' regions where the percentage of GC is higher. This uneven distribution of GC content in *P. falciparum* is predominantly due to the high GC content of the subtelomeric *var* genes. Exactly defining where the subtelomeric regions end and where the core begins has remained debatable, but recent work using multiple long-read assemblies has attempted to resolve this using gene orthologies[255]. The current 3D7 genome annotation lists 5,432 genes, with 36% (1,964/5,432) of them being annotated as unknown function, reflecting the amount of *Plasmodium* specific biology that we still do not fully understand. However, this compares with 60% of genes having unknown function upon the initial publication of the genome in 2002[85], showing to an extent the progress that has been made in understanding *Plasmodium* genetics over the last 15 years.

0.2.3 Sequencing Other *Plasmodium* Species

While *P. falciparum* was the first human malaria parasite species to be sequenced, *P. vivax* and *P. knowlesi* have also had reference genome sequences assembled and annotated for them over time[51,266]. The first *P. vivax* reference genome sequence was published in 2008, originating from a strain (termed Sal1) isolated in 1972 from a patient in El Salvador[51]. While a breakthrough at the time, the *P. vivax* Sal1 reference genome sequence was less contiguous than the *P. falciparum* 3D7 genome, with thousands of contigs that could not be assigned to chromosomes and with the 14 core chromosome regions being split across 30 contigs[51,20]. Additionally, the gene annotation of the SAL1 genome was not manually curated by the original authors and therefore over time began to lag in accuracy to other curated *Plasmodium* reference genomes. It was in 2016, that a new *P. vivax* reference genome was published (PvP01) which managed to consolidate the 14 chromosomes to 14 contigs, reduced the

amount of unassigned contigs to 226, and brought the gene annotation up to date with the other reference genome sequences[20]. The *P. vivax* PvP01 assembly spans 29Mb of sequence and has a GC content of 39.8%, which is significantly higher than *P. falciparum*. *P. vivax* also exhibits an uneven GC content between the core and the subtelomeres, though in *P. vivax* this does not appear to be linked to a particular gene family. Conversely to *P. falciparum*, *P. vivax* has a higher level of AT in the subtelomeres and a lower level in the core regions of the chromosomes[51], referred to as an 'isochore' structure. The subtelomeres in *P. vivax* are much larger than those in *P. falciparum* and mostly account for the difference in genome size between the two assemblies (29.0Mb vs. 23.3Mb)[20]. Due to the larger genome size, the *P. vivax* PvP01 assembly with 6,642 genes has many more genes than the *P. falciparum* 3D7 reference. Most of these additional genes are located in the subtelomeric regions of the PvP01 reference and belong to large multigene families. While there are major differences in the structure and gene content of the subtelomeric regions, the core regions of the *P. vivax* genome share a large amount of synteny with those in *P. falciparum*, with most core genes being positionally conserved[51].

Shortly after the publication of the *P. vivax* Sal1 reference genome, the genome sequence of the zoonotic *P. knowlesi* was published[266]. The 23.5Mb genome of *P. knowlesi* H strain was slightly less discontiguous than the original *P. vivax* Sal1 assembly, but still consisted of 715 contigs, 511 of which couldn't be assigned to any of the 14 chromosomes[266]. With 5,188 genes, *P. knowlesi* has fewer genes than either *P. falciparum* and *P. vivax*. The *P. knowlesi* genome also differs from the latter two by having multiple intrachromosomal regions with high GC content that contain tandemly repeating telomeric sequences (heptad sequence GGGTT[T/C]A)[266]. Besides these species-specific regions, much of the *P. knowlesi* genome was again highly syntenic to the other sequenced human malaria

parasite species[266]. Very recently, an updated *P. knowlesi* genome was published using PacBio long reads, reducing the number of contigs down to 28[180].

Up until the beginning of this project, neither the genomes of *P. malariae* nor *P. ovale* had been sequenced, leaving the number of sequenced human malaria parasite species at three out of five. In 2016 however, draft genomes for both species (including the two *P. ovale* subspecies) were published by Ansari et al. [10]. An analysis of these draft genomes is presented in Chapter 1. Finally, besides the human-infective *Plasmodium* species, a number of genome sequences have been published for rodent-infective, primate-infective, and avian-infective species.

Rodent-infective malaria parasite species have long been used experimentally as models of human malaria, with a number of these species having been adapted to passaging through laboratory mice[73]. Around the same time as the completion of the *P. falciparum* genome project, the complete genome of the rodent-infective *P. yoelii* was published[52]. Other rodent malaria genome sequences followed for *P. berghei* and *P. chabaudi*[131], with improved versions published in 2014[256]. All these rodent malaria genomes are small in size, with many below 20Mb, and have GC contents of ≈30%[256]. The genomes are extremely collinear in the core regions and nucleotide-sequence identity between the genomes is very high at ≈90%[171].

Other primate-infective *Plasmodium* species have been sequenced in order to get a better understanding of the evolutionary history of human malaria parasites and to see how the latter have adapted specifically to infect humans. The first primate-infective species to be sequenced was *P. cynomolgi*, which infects simian monkeys, and which appears to resemble *P. vivax* closely both phenotypically and genetically[341]. Following this, the chimpanzee-infective *P. reichenowi* was sequenced[259]. Due to being very similar morphologically, it was long thought that *P. reichenowi* was the closest extant rel-

ative to *P. falciparum*, which was supported by the genome sequences of the two species being more alike each other than to any other sequenced *Plasmodium* species, forming the so-called '*Laverania*' subgenus . Recently however, genome sequences for a number of *Plasmodium* species in the *Laverania* subgenus have been published, shedding light on the *Plasmodium* phylogeny (see below)[258], concluding that *P. falciparum* likely originated from a gorilla-infective species called *P. praefalciparum*[200]. The genome sequences of these *Laverania* species resemble that of *P. falciparum* in terms of gene content and structure[258], clearly distinguishing them from the species in the *P. vivax* clade, as well as the rodent malaria parasites.

0.2.4 Population Genomics of Malaria

As the cost of sequencing continued dropping and sample collection methodologies improved, the number of malaria parasite whole-genome sequences kept increasing rapidly, enabling the study of genetic variation across time and space. This enables for instance the identification of gene flow between populations, tracking of changes in parasite population due to control interventions, and pinpointing of drug resistance mutations when they emerge. The first study to look at multiple different *P. falciparum* genome sequences examined 16 samples[359], in 2012 a study looked at 227 samples[209], while the current Pf3k dataset, an international collaboration to sequence *P. falciparum* samples from across the globe, includes 2512 *P. falciparum* whole-genome sequences (www.malariagen.net). For *P. vivax*, the first study to look at multiple whole genome sequences was only published in 2012 and looked at four sequences[239], while more recent studies have included hundreds of samples[147,274]. Finally, even for the less studied *P. knowlesi*, multiple whole genome sequences have been analysed[281].

In order to analyse these large datasets, software and numerous specialised methodologies have

been developed. The initial step is to map all samples against a reference genome sequence, thereby producing an alignment of the sample reads to the reference genome, *ie.* determining which region of the genome each sequencing read corresponds to. Most mapping software is based on either hashing algorithms, such as MAQ [193] and Stampy [205], or on the so-called 'Burrows-Wheeler transform' data compression algorithm [242], including Bowtie [179] or BWA [191]. When running these tools, parameters are set to determine the leniency with which the reads are mapped, *ie.* how different the read can be from the reference sequence whilst still being aligned. The mapping software then provides metrics such as the proportion of reads that mapped and whether they were properly paired (in case of paired sequences), which can be used as metric for the quality of the sample and sequencing data. The type of information that can be garnered from viewing aligned reads includes getting an idea about the general coverage of the genome, *ie.* whether all the regions of the genome are covered and whether the coverage is evenly distributed, and also highlights the general level of similarity of the sample and the reference genome by pinpointing discrepancies between the sequences such as sample contamination or SNPs (figure 7). Sample reads that do not map to the reference genome can also be of interest depending on the reason they didn't map, such as in cases of highly polymorphic regions, *ie.* too many differences between sample reads and reference genome, or if the sample has novel genetic elements not found in the reference sequence, as is often the case with antigenic gene families. Based on the manual inspection of read mapping, samples may be excluded from further analysis if they are deemed to be of low quality, a judgment that can be made based on overall coverage or evidence of contamination for instance.

Once the final sample set is determined, genotypes, and consequently SNPs, need to be called from the mapped reads [242]. The fundamental concept behind SNP-callers is to look at the allele distribution

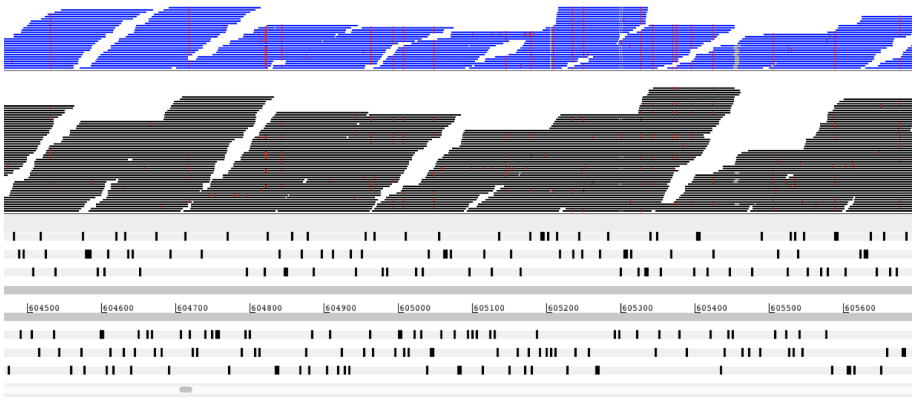


Figure 7: Showing an example Artemis [300] view, with reads from two samples (black and blue respectively) mapped to a reference genome sequence (bottom). While both samples have even coverage, the black sample appears to have higher levels of coverage than the blue sample (indicated by the height of the bars, ie. how many reads are stacked on top of each other). Red marks indicate discrepancies between the sequence reads and the reference sequence. Consistent marks that are present in all reads covering a specific part of the reference sequence are likely SNPs, examples of which can be seen in the blue sample. Inconsistent red marks are potentially sequencing errors.

in the sequenced reads for a specific base call and to calculate the probability of that position being a specific allele. By using multiple samples, such as for GATK [215], the additional information of the general distribution of alleles for that base position can be used, whereas linkage-disequilibrium gives another independent source of information for certain SNPs that are difficult to call [242]. Following SNP-calling, SNPs are filtered both by sample, *ie.* certain samples may have a high proportion of missing calls, and by base position, as certain SNPs may be in highly polymorphic regions and the SNP calls may therefore be unreliable [242]. SNP filtering can often be done using in-built functions in the SNP calling softwares, such as in GATK where best practices for filtering SNPs have been published [215]. The filtered SNP set can then be used for a number of analyses, such as looking at nucleotide diversity, positive or negative selection acting on genes, or looking at SNPs that segregate between populations for example. Such analyses have already yielded significant insights into the genetic basis of malaria.

0.3 Insights Gained from Malaria Genomics

0.3.1 Subtelomeric Gene Families

One of the early insights from sequencing different *Plasmodium* species was the observation that each newly sequenced genome appeared to contain an abundance of genes of unknown function that were specific to that species. These large species-specific gene expansions often cluster into 'gene families', meaning that the genes in the same family share higher levels of nucleotide similarity and gene structure with each other than with other genes in the genome [157,173]. The similarity in sequence and gene structure of genes in the same gene family suggests that they likely originated from an ancestral expansion event and subsequently began to diversify [157]. In most *Plasmodium* species, these gene families are often restricted to the highly recombinogenic subtelomeric regions of the genome [309], though there are exceptions such as instances of chromosome-internal gene clusters in *P. knowlesi* and *P. falciparum*. Between individuals of the same species, these subtelomeric gene families vary tremendously both in number and in nucleotide similarity, due to significant amounts of recombination in the subtelomeric regions. This high amount of variation in sequence, genomic location, gene presence/absence, and, paradoxically, non-specificity due to high levels of nucleotide similarity between gene family members in the same genome, make these genes incredibly difficult to study using conventional population genetics approaches [25]. Being able to confidently map sequencing reads to these genes is rare and calling reliable SNPs from these alignments is consequently very difficult. In order to be able to make sense of these gene families, they are often assembled *de novo* from either unmapped reads or from reads mapping to the subtelomeric regions and the resulting arsenal of genes can then be compared between individuals using comparative genomics approaches [173,20].

Even before the sequencing of the first *P. falciparum* genome, research into the so-called *var* genes and their role in antigenic variation was on-going [336,26]. We now know that each *P. falciparum* genome contains approximately 60 *var* genes and a number of *var* pseudogenes. Close evolutionary relatives of *P. falciparum*, such as *P. reichenowi* and other *Laverania* species, also have *var* gene repertoires, but none of the other human-infective species does. The *var* genes encode the *P. falciparum* erythrocyte membrane proteins (PfEMP1), which are transported to the red blood cell surface [361]. At any point in time only one *var* gene is expressed [310], with the switching between *vars* thought to potentially be mediated by a highly conserved *var* gene known as *var2csa* [352]. It is believed that by switching between different *var* genes, the parasite is able to evade the host immune system [361]. The PfEMP1 proteins are thought to bind to a number of host proteins, such as complement receptor 1 (CR1) [299] and heparan sulfate [58], and that this binding may lead to rosetting [66], *ie.* agglutination of infected red blood cells, which in turn may result in cases of cerebral malaria [50]. In addition to cerebral malaria, expression of *var2csa* leads to PfEMP1 binding to chondroitin sulfate A (CSA) in the placenta of pregnant women [65], resulting in placental malaria [304]. Due to the involvement of *var* genes in causing severe malaria, they have been extensively studied.

In addition to *var* genes, the *P. falciparum* genome also contains members of other gene families, including *etramp* genes, *phist* genes, and *rif* genes. The latter are now thought to be part of a large gene family present in all sequenced *Plasmodium* species, the *pir* (*Plasmodium* interspersed repeat) genes [157,76]. The *pir* genes also have species specific names, such as *vir* in *P. vivax*, *kir* in *P. knowlesi*, and *cir* in *P. chabaudi* for example. The number of *pir* genes varies significantly between species, with *P. relictum* (an avian malaria species) for instance only having 4 copies [38], while *P. falciparum* has 227, and the newly assembled *P. vivax* P01 genome appears to have an astounding 1,212 *pir* genes [20], in line

with the 1,373 *pirs* in the closely related *P. cynomolgi* genome[269]. While the *pir* genes between species appear to be related to each other, and likely have a single common ancestor, the *pir* repertoire of each species is distinct, suggesting that they expanded in a species specific manner[76]. The exact function of *pir* genes is not known, though it was originally thought that the large number of genes and high variability between the genes may imply a role in antigenic variation, similar to *var* genes. Studies in *P. vivax* found a role of *pir*-encoded proteins binding to endothelial cells[30], while studies in *P. falciparum* suggested an involvement in rosetting through binding to red blood cells belonging to blood group A[121]. Work in mouse models with *P. chabaudi* suggests that *pir* genes may be involved in virulence and in establishing chronic infections depending on the type of *pir* genes being expressed[329,44]. While their biological function is still under investigation, the sheer abundance of *pir* genes across all sequenced *Plasmodium* species perhaps indicates important, yet to be discovered, roles.

0.3.2 *Plasmodium* Phylogenetics

Disentangling the evolutionary relationship between *Plasmodium* species has been a difficult challenge and the inferred evolutionary tree has been revised numerous times[293]. In part, this is due to the complexity of *Plasmodium* genetics, where different parts of the genome tell different evolutionary stories[133], as well as the large differences in GC content between the species skewing relationship estimates[112]. Early phylogenies often relied on small parts of the genome, such as 18S rRNA or the *cytochrome B* locus of the mitochondrion for example[293]. This, together with the fact that many key *Plasmodium* species hadn't been discovered yet, led to a number of conclusions that have now had to be revised[293]. Some of the biggest questions in *Plasmodium* phylogenetics relate to the evolutionary origin of the human-infective species, with a lot of work having been done on both *P. falciparum* and

P. vivax.

The hypothesis of the origin of *P. falciparum* has been revised a number of times. It was initially thought that the particularly high virulence of *P. falciparum* indicated that it had jumped into humans very recently from birds[362]. This bird-origin hypothesis was however rejected when a number of primate-infective *Plasmodium* species were discovered and described as *P. falciparum*-like (*P. reichenowi*), *P. malariae*-like (*P. rhodaini*), and *P. vivax*-like (*P. schwetzi*), due to their morphological similarities[291]. The rRNA sequencing of the chimpanzee-infective *P. reichenowi*, showed the close relationship of the species to *P. falciparum* and it was suggested that the two species shared a common ancestor up to the event of humans diverging from chimpanzees, at which point they began to co-evolve in parallel with their newly diverged hosts[97]. However, this hypothesis therefore suggested that *P. falciparum* had co-evolved with humans for millions of years, which was at odds with its highly virulent nature. It was only when additional species of *Plasmodium* were discovered in other primates that the phylogeny around *P. falciparum* became clear[291]. Extensive sampling of primate-infective species revealed that *P. falciparum* likely originated from a Gorilla-infective species now termed *P. praefalciparum*[200]. Further sequencing confirmed the close similarity of *P. praefalciparum* to *P. falciparum*[258,181] and highlighted a number of key genetic changes that may have enabled the host switch from Gorillas to humans[338]. Furthermore, the close genetic similarity of the two species suggests that the host switch occurred relatively recently, with estimates ranging from 10,000[338] to 50,000[258] years, which is in line with the high virulence of *P. falciparum*.

While the origin of *P. falciparum* is now generally accepted to be a host switch from Gorillas to humans with the closest extant relative being *P. praefalciparum*[258], the evolutionary origin of *P. vivax* is still contested. Specifically, the debate centers around whether *P. vivax* has an African or an Asian

origin. For many years, the consensus was that *P. vivax* likely originated in Southeast Asia from a cross-species transmission event of a macaque parasite[229,239]. This hypothesis was supported by the observation that the closest known relative of *P. vivax* was for a long time the macaque-infective *P. cynomolgi*[341] and that most of the other close relatives of *P. vivax* are other simian-infective species such as *P. simiovale* or *P. knowlesi*[204]. This out-of-Asia hypothesis however cannot account for two important facts. The first is that natural resistance to *P. vivax* is widespread throughout Africa in the form of Duffy-negativity[54], *ie. P. vivax* struggles to infect erythrocytes lacking the duffy antigen. The second is that humans only arrived in Asia 60,000 years ago[216] while *P. vivax* likely diverged from macaque-infective species much longer ago than that[229,239]. The recent discovery of great apes infected with *P. vivax* throughout Africa has further convoluted this out-of-Asia story by necessitating an importation of *P. vivax* from Asia to Africa to account for this observation[286]. This discovery of *P. vivax* in African great apes has however strengthened the out-of-Africa hypothesis, proposing that *P. vivax* originated from a host-switch from African great apes into humans and then subsequently spread to Asia, before essentially disappearing from humans in Africa due to humans evolving Duffy-negativity[75]. This hypothesis is further supported by the recent sequencing of a chimpanzee-infective *Plasmodium* species that now appears to be the closest relative of *P. vivax*[204], known as *P. vivax*-like. The main problem with the out-of-Africa proposal is the phylogenetic placement of *P. vivax* and its chimpanzee-infective relative within the clade of Asian monkey-infective species. The question of out-of-Asia versus out-of-Africa is therefore still ongoing.

While the evolutionary origin of *P. falciparum* and *P. vivax* has been extensively studied and debated, it is not surprising that little is known about the evolutionary origin of *P. malariae* and *P. ovale*. As both species only result in relatively benign forms of malaria, it is often believed that both species

have been infecting humans for a very long time. This is further supported by the ability of *P. malariae* to persist in its host for many decades[318], suggesting that it has exquisitely adapted to humans over a long period of time. Little is known about their host range, however it is believed that the New World monkey-infective *P. brasilianum* could potentially be *P. malariae*. The limited phylogenetic studies that have been performed with *P. malariae* and *P. ovale* place both of them closer to *P. vivax* than to *P. falciparum*[112,204,14], though the exact placement of both species is still contested. Both species usually form outgroups to *P. vivax* and the Asian primate-monkey infective species, though published phylogenies differ in terms of placing *P. malariae*[112] or *P. ovale*[204] as the furthest outgroup. A study based on the apicoplast, placed *P. ovale* as a sister taxa to the rodent-infective species and *P. malariae* as an outgroup to that clade, suggesting a potential host switch from humans to rodents[14]. These phylogenies all have in common that both *P. ovale* and *P. malariae* significantly differ from other studied *Plasmodium* species and it has therefore been difficult to accurately determine their phylogenetic relationship to these. Full genome information will enable both species to be more accurately placed within a phylogeny.

0.3.3 Genetics of Antimalarial Drug Resistance

During the GMEP, chloroquine resistance emerged and spread across the globe. Since then, *P. falciparum* has managed to evolve resistance to almost every drug that has been thrown at it[36] (figure 8). Using advances in sequencing, it has become possible to study the genetic basis of chloroquine resistance as well as the genetic basis of antimalarial resistance to other drugs deployed since then. The fundamental idea in identifying mutations or other genetic variants that may be involved in providing resistance to specific drugs essentially consists of comparing sensitive and resistant parasites and then

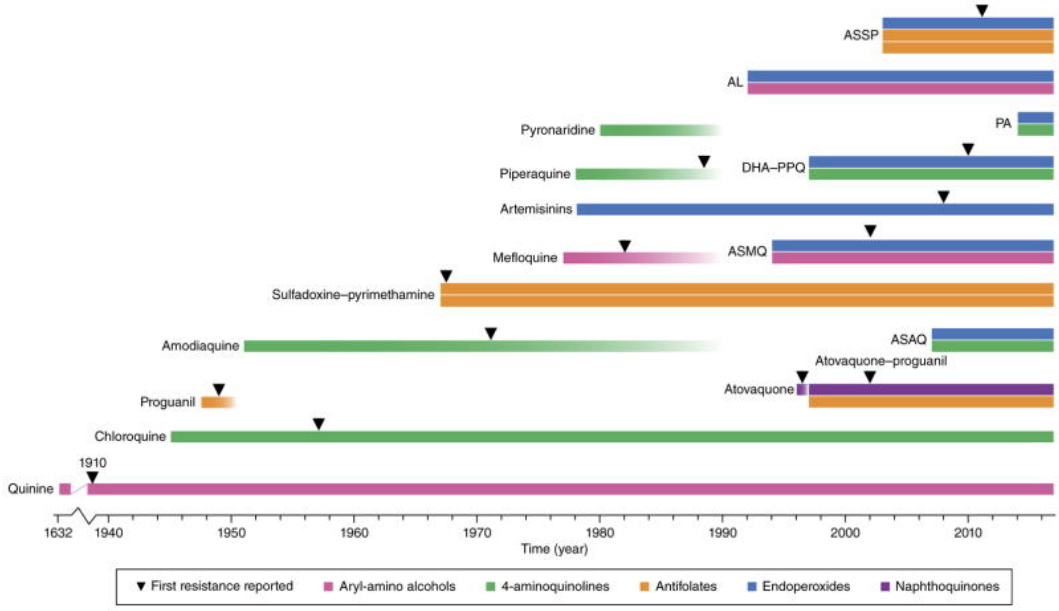


Figure 8: Showing the timeline of different antimalarial drugs being deployed and the timing of the first resistance to the drugs being detected. Figure reproduced from Blasco et al. [36].

seeing what differs between them. This procedure therefore consists of two important parts, the first being the process of determining the level of resistance of the parasites to be studied, and the second being the method used to compare these parasites.

For the first part, there are generally two different ways of assessing the level of resistance of parasites. The first consists of using treatment outcomes, where parasites are sampled before treatment and after treatment. Those that are sampled after treatment are likely to be the resistant ones, while those before treatment are a combination of resistant and sensitive parasites. A comparison between these groups of parasites would then pinpoint certain genetic variants that are either significantly enriched or significantly depleted in the resistant population compared to the baseline population. A more common method of ascertaining the level of resistance of parasites is to perform either *in vitro* or *ex vivo* drug testing to measure their 50% inhibitory concentration (IC_{50}) value[3,5,8]. These IC_{50}

values are a quantitative estimate of how much of a drug is needed to kill 50% of the parasites and are estimated by fitting a dose response curve after exposing parasites to a range of different drug concentrations. Compared to the binary outcome of the simple before and after counting method, IC_{50} values are quantitative estimates and therefore provide a higher level of resolution, though they can be laborious to measure in the first place. For certain drugs, such as artemisinin, IC_{50} values have not been accurately associated with the actual clinical response to the drug, and alternative measures have been developed such as the ring-stage survival assay (RSA)[13].

Once estimates of drug resistance have been obtained for studied samples, analytical methods need to be employed to pinpoint the genetic variants that are most strongly associated with the observed drug phenotype. An important method that has only recently become amenable to be used in *Plasmodium* due to the large number of sample phenotypes and whole genome sequences that are required for statistical confidence, is performing genome-wide association studies (GWAS)[275]. GWAS studies perform statistical association tests for every SNP in the genome with the tested phenotype[358]. Due to the large number of tests that this entails (many thousands depending on the amount of SNPs), the number of false positives, *ie.* associations that are significant by chance, will often times outnumber the true positives[154]. To account for this, GWAS studies adopt multiple testing corrections and thereby require associations to be exceedingly significant to reach 'genome wide significance'[358]. To achieve this, either the underlying effect of a variant on a phenotype has to be very large, or a significant number of samples have to be examined to provide sufficient statistical confidence[145]. GWAS studies can also be obfuscated by other factors, such as underlying population structure, which can be a major problem in *Plasmodium*[221], and choice of study population, where the genetic basis of resistance may differ between populations[358]. Finally, not all resistance phenotypes are associated with

Table 1: Antimalarial Drugs and their Genetic Markers

Antimalarial Drug	Genetic Markers	References
Chloroquine	*crt*,*mdr1*, *mrp*	364,81,105,321,158,288,354
Mefloquine	*mdr1* (SNPs & CNV), *mspdbl2*	370,89,282,319,230,353,320,354
Piperaquine	*plasmepsin 2/3* (CNV), *exonuclease*, *mdr1* (SNPs & CNV)	6,92,354,8,372
Artemisinin	*kelch 13*, *mdr1* (SNPs & CNV)	89,319,13,221,213,334,354
Lumefantrine	*mdr1* (SNPs & CNV), *mspdbl2*	319,353,354

changes in SNP frequencies, other underlying resistance mechanisms may be responsible, such as gene copy number amplifications or epigenetic changes, though GWAS analyses can be adapted for those purposes.

Before the advent of GWAS, drug resistance genes in *P. falciparum* were initially identified using homology, such as for the *multidrug resistance protein 1* (*mdr1*) gene involved in mefloquine resistance[105,369] or the *dihydropteroate synthase* (*dhps*)[349] and *dihydrofolate reductase-thymidylate synthase* (*dhfr*)[277,276] genes that are respectively involved in sulfadoxine and pyrimethamine resistance. An alternative method was by using genetic linkage information following selection experiments[358], leading to the identification of the *chloroquine resistance transporter* (*crt*) gene responsible for chloroquine resistance[364,81]. These initial findings have subsequently been recapitulated by GWAS studies[267,358], with *crt* and *mdr1* being significantly associated with resistance to a number of antimalarial drugs, including chloroquine and mefloquine[230]. Other GWAS studies have subsequently identified novel genes associated with resistance to amodiaquine, quinazoline, quinine, halofantrine, mefloquine, lumefantrine, piperaquine and artemisinin[365,353,8,372,57,221,342]. See table 1 for a list of selected antimalarial drugs and the genes with which they have been associated[358].

Pinpointing the genetic basis of artemisinin resistance and piperaquine resistance have been important breakthroughs in handling the current crisis of multidrug resistance in Southeast Asia. Artemisinin

resistance, as characterised by a reduction in the speed of parasite clearance, has been linked to mutations in the propeller domain of the *kelch 13* gene[219,221,13], and piperaquine resistance is strongly associated with a copy number amplification of the *plasmepsin 2* and *plasmepsin 3* genes[8,372] (as well as with a SNP in an *exonuclease* gene, which is in strong linkage with the *plasmepsin 2/3* CNV[8]). While artemisinin resistance has been shown to have emerged multiple times in Southeast Asia[342], one particular lineage, KEL1, outperformed others and eventually combined with the PLA1 lineage, associated with piperaquine resistance[9]. This multidrug resistant KEL1/PLA1 lineage is the main agent of the current outbreak of multidrug resistance in Southeast Asia[9,150], having spread from Western Cambodia to other countries[149], including Thailand, Laos, and Vietnam[6,149,307,346]. Knowing the genetic basis of these multidrug-resistant *P. falciparum* parasites now enables us to better track them and may even help us in deciding how to best fight them next.

0.4 Thesis Overview

Malaria is a complex disease made up of a multitude of different interlinking factors. Perspectives on the field of malaria research will vary significantly from person to person, ranging from parasitologists to entomologists, clinicians to bioinformaticians, and even from sociologists to economists. In this thesis, I have attempted to outline and explore a number of key gaps in our understanding of malaria, specifically from the perspective of a computational parasitologist with an interest in evolutionary biology. As a result of this narrow focus, little attention is given to otherwise very important aspects of malaria research, such as vector control or vaccine development. Throughout this thesis, the main leading question that guides the work is: 'What can genomics tell us about human malaria parasites?' (figure 9).

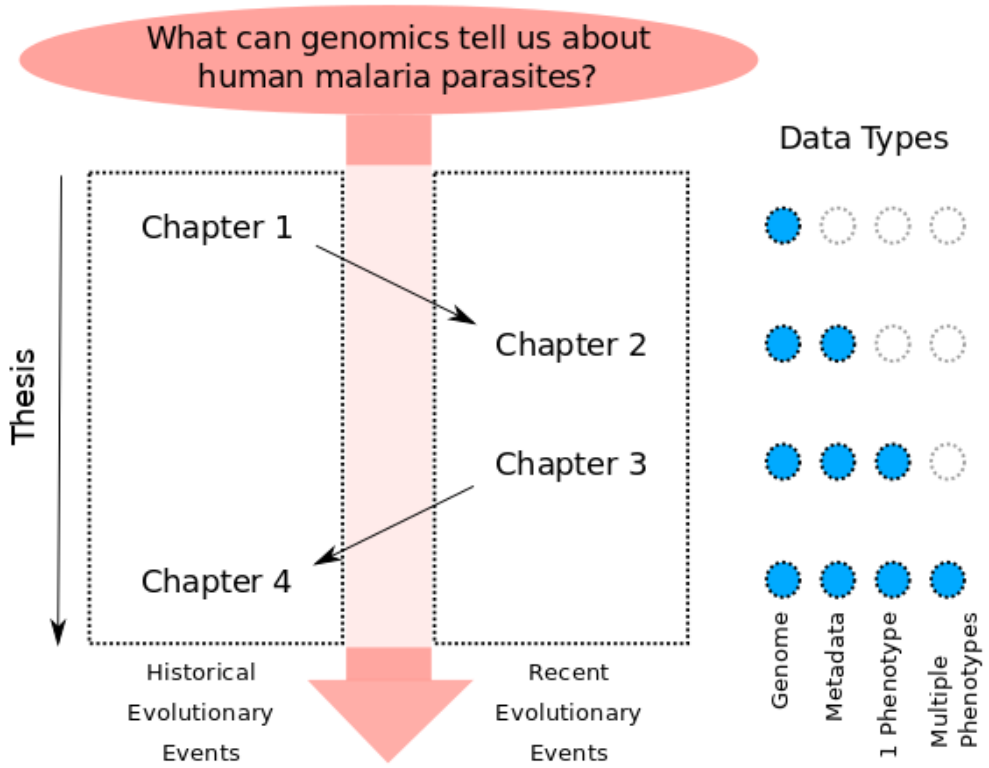


Figure 9: Showing an overview of the thesis structure, with the guiding question in red.

As discussed above, genomics has already taught us a lot about the biology of the malaria parasite, however there is still a lot to be learned. Genomics has the ability to shed light both on the distant past, through constructing phylogenies and identifying evolutionary adaptations, as well as on recent events, by tracking outbreaks of drug resistance and diagnosing hospital infections. While there may be a temporal disconnect between the timescales of these applications of genomics, they do intersect and influence each other. In this thesis, I will attempt to delve into both the ancient and the recent, and showcase the intersections between them. To do this, I will identify important problems in both realms, explorations of which I will then apply in the other (figure 9).

One of the most surprising gaps in our understanding of the evolutionary history of human malaria

parasites is the lack of reference genome sequences for two of the five human malaria parasites, *P. malariae* and *P. ovale*. We know little about their phylogenetic relationship to other *Plasmodium* parasites and even less about the genes that phenotypically distinguish these species from the more virulent human-infective ones. With the rapid advances in sequencing and sample preparation over the last two decades, obtaining sufficient sequencing data to assemble reference genomes for both species should now be possible. Not only would these genome sequences enable us to answer important evolutionary questions about these species, but it would also open up the possibility of studying recent cases of treatment failure or co-infections in these species from a genomic perspective.

On the scale of recent evolutionary events, the current rise of multidrug resistance in Southeast Asia necessitates a genomic approach to understanding and tracking the problem. While KEL1/PLA1 parasites have become resistant to both artemisinin derivatives and piperaquine, little is known in terms of their specific response to other antimalarial drugs, such as mefloquine. Analysing whole genome sequences of KEL1/PLA1 parasites with associated mefloquine IC_{50} values using a GWAS approach, will answer important questions relating to the individual response of KEL1/PLA parasites to mefloquine and the likely sustainability of mefloquine in the field longer term. Notably, the large number of whole genome sequences collected as part of the Pf6 dataset over the last two decades will then enable us to take a historical perspective of multidrug resistance in the region and beyond.

One additional leitfaden throughout this thesis is that of layering data, where, at the beginning of the thesis, I look solely at genome sequences and by the end of it, I amalgamate data ranging from genotype data to metadata, over to clinical phenotypes (figure 9). In Chapter 1, I assemble reference genome sequences for both *P. malariae* and *P. ovale*, analysing their gene content and relating it to their unique biology. I also use additional draft genome assemblies to confirm whether *P. ovale* con-

sists of two species. In Chapter 2, clinical metadata is layered on to the genome sequences in order to characterize a case of clinical recrudescence of a *P. malariae* infection. In Chapter 3, phenotype data on mefloquine susceptibility is used together with *P. falciparum* whole genome sequences and metadata to perform a GWAS of mefloquine resistance in Southeast Asia. Finally, in Chapter 4, phenotype data for multiple drugs is harnessed, including chloroquine, artemisinin, mefloquine and piperaquine, in order to understand the genetic basis of multidrug resistance. At each level, I attempt to identify new biology that both expands our understanding of human malaria in general and sheds light on the specifics of antimalarial drug resistance.

Our castle's strength will laugh a siege to scorn. Here let them lie till famine and the ague eat them up.

William Shakespeare, Macbeth, 1606 CE

1

Completing the Set of Human Malaria Parasite Genomes

1.1 Abstract

Despite the huge international endeavor to understand the genomic basis of malaria biology, there remains a lack of information about two human-infective species: *Plasmodium malariae* and *P. ovale*. The former is prevalent across all malaria endemic regions and able to recrudesce decades after the

initial infection. The latter is a dormant stage hypnozoite-forming species, similar to *P. vivax*. In this chapter, I present the newly assembled reference genomes of both species, thereby completing the set of all human-infective *Plasmodium* species. I show that the *P. malariae* genome is markedly different to other *Plasmodium* genomes and relate this to its unique biology. Using additional draft genome assemblies, I confirm that *P. ovale* consists of two species that appear to have diverged millions of years ago. These genome sequences now provide a new resource to study the genetic basis of human-infectivity in *Plasmodium* species and open up otherwise impossible opportunities for developing diagnostics, drugs and vaccines against these neglected species of malaria parasites.

1.2 Introduction

All known human malaria species were described in the early 20th Century, with *Plasmodium malariae* and *P. ovale* being recognized as distinct species from *P. falciparum*, *P. vivax*, and *P. knowlesi* [166]. Reference genomes have now been published for the latter three [115,51,266], with the extent of human infections caused by *P. knowlesi* having only been recognized decades after initial discovery [60]. Analysis of these reference genomes has revealed the genomic basis of key biological processes, including virulence [76], invasion [69], and antigenic variation [228]. The lack of whole genome sequences for *P. malariae* and *P. ovale* has meant that little progress has been made in the understanding of molecular genetics in these important species.

Infections with *P. malariae* and *P. ovale* are frequently asymptomatic [298] and often have parasitaemia levels undetectable by light microscopy [83], making their study in human populations difficult and potentially thwarting efforts to eliminate them and declare any regions 'malaria free' [40]. This lack of knowledge is especially worrying because the two species are distributed widely across

all malaria-endemic areas of the world[67,68] (figure 1.1 a). *P. malariae* and *P. ovale* frequently occur as co-infections with the two more common species, *P. falciparum* and *P. vivax*, and can be present in up to 5% of all clinical malaria cases[298]. This equates to roughly 30 million annual clinical cases. *P. malariae* infections can lead to lethal renal complications[178] and can recrudesce after decades[318], further increasing their socioeconomic costs.

Unraveling the mechanisms that enable *P. malariae* to persist in the host for decades is critical for a more general understanding of chronicity in malaria. The genome sequence of *P. ovale*, the other hypnozoite-forming species, will facilitate the search for conserved hypnozoite genes and will conclusively show whether *P. ovale* consists of two cryptic subspecies, as recently suggested[339]. Finally, the genetic basis of human-infectivity in malaria parasites can only be fully understood by having access to the genome sequences of all human-infective species.

Here I present the genome sequences of both these species, including the two recently described[339] subspecies of *P. ovale* (*P. o. curtisi* and *P. o. wallikeri*). I update the phylogeny of the *Plasmodium* genus using whole genome information, and describe novel genetic adaptations underlying their unique biology. Using whole genome sequencing of additional *P. malariae* (including two obtained from chimpanzees, referred to as *P. malariae*-like) and *P. ovale* samples, I describe their genetic variation, as well as identify genes that are under selection. The data presented here provide the community with an essential foundation for further research efforts into these neglected species and into understanding the evolution of the *Plasmodium* genus as a whole.

All methods used for this chapter are detailed in Appendix A. This project was a large collaborative effort with sample collection and sequencing having been performed by a large number of collaborators. All data analyses presented in this chapter are my own work unless specifically noted otherwise.

1.3 *Plasmodium* Co-Infections

Obtaining *P. malariae* and *P. ovale* DNA has historically been difficult due to the low level of parasitaemia in natural human infections. Using a novel method based on identifying species-discriminating mitochondrial SNPs I developed (Appendix A), *P. malariae* and *P. ovale*were found in approximately 2% of all *P. falciparum* clinical infections from the globally sampled Pf3K project (www.malariagen.net) (figure 1.1 a) (table 1.1), compared to 4% being co-infections with *P. vivax*. A number of infections containing three species were also identified. These *P. malariae* and *P. ovale* co-infections are in addition to the larger number of mono-infections that they cause, which are frequently missed due to difficulties in confirming a species diagnosis. I used the two *P. ovale* co-infections with the highest number of sequencing reads to perform *de novo* genome assemblies.

1.4 Genome Assemblies

A 33.6 megabase (Mb) reference genome of *P. malariae* was produced from clinically isolated parasites and sequenced using Pacific BioSciences long-read sequencing technology. The assembled sequence comprised 14 super-contigs representing the 14 chromosomes, with 6 chromosome ends extending into telomeres, and a further 47 unassigned subtelomeric contigs containing an additional 11 telomeric sequences (table 1.2). Using existing Illumina sequence data from two patients primarily infected with *P. falciparum*, sequencing reads were extracted and assembled into 33.5Mb genomes for both *P. o. curtisi* and *P. o. wallikeri*, each assembly comprising fewer than 800 scaffolds. The genomes are significantly larger than previously sequenced *Plasmodium* species and, like *P. vivax*, have isochore structures with a higher AT content in the subtelomeres. In addition, a *P. malariae*-like genome

Table 1.1: Samples positive for different Plasmodium species in the Pf3k dataset

Country	Total Samples	*P. falciparum*	*P. vivax*	*P. malariae*	*P. ovale*	*P. knowlesi*
The Gambia	65	65	0	0	0	0
Guinea	100	100	0	7	3	0
Thailand	148	148	11	0	0	0
Ghana	617	617	5	12	9	0
Cambodia	570	570	50	0	0	0
Mali	96	96	0	1	0	0
Bangladesh	50	50	4	0	0	0
Malawi	369	369	4	4	4	0
Vietnam	97	97	16	0	0	0
Myanmar	60	60	7	0	0	0
Laos	85	85	4	0	2	0
DR Congo	113	113	1	2	1	0
Nigeria	5	5	0	0	0	0
Senegal	137	137	0	0	1	0
Global	2512	2512	102	26	19	0

The first column shows country of origin for the different samples, with the second column showing the total number of samples collected in that country. The following five columns show the number of these samples that are positive for the different *Plasmodium* species. All samples are positive for *P. falciparum*, which is expected because all the samples were initially identified as *P. falciparum* infections. I did not see any samples positive for *P. knowlesi*.

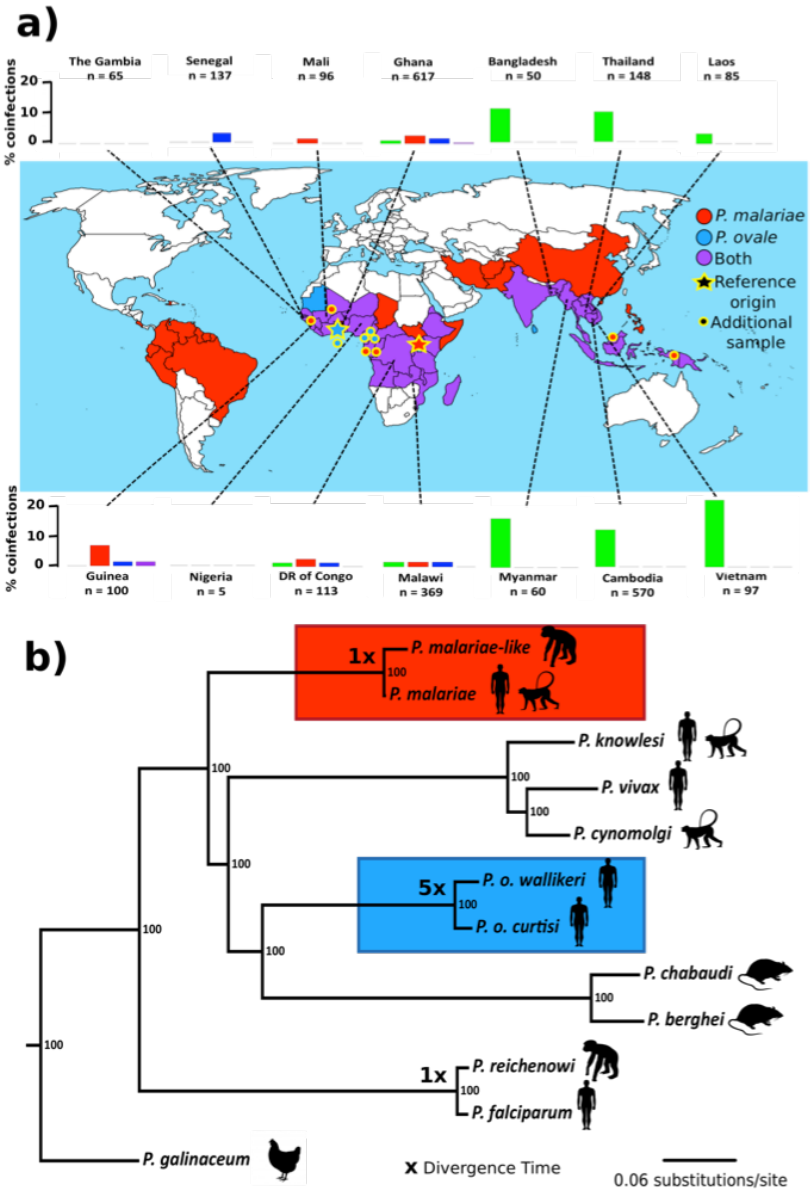


Figure 1.1: Prevalence and phylogenetic relationship of P. malariae and P. ovale. a) World map showing presence and absence of P. malariae (Red), P. ovale (Blue) or both (Purple) by country based on a literature review. Barplots show proportion of P. falciparum infections with co-infections of P. malariae (Red), P. ovale (Blue), P. vivax (Green), or two species (Purple) based on the Pf3K dataset. Stars indicate origin of sample used for reference genome assembly and points show additional samples used. b) Maximum likelihood phylogenetic tree of the Plasmodium genus, showing the P. malariae clade (Red) and the P. ovale clade (Blue) together with the divergence times of the species as calibrated to the P. falciparum/P. reichenowi split (x). Silhouettes show infectivity of the different species. Values within branching points indicate bootstrap values as calculated using RAxML [331].

Table 1.2: Comparison of genome features of all human-infective Plasmodium species and P. malariae-like

Feature	*P. falciparum*	*P. knowlesi*	*P. vivax*	*P. ovale curtisi*	*P. ovale wallikeri*	*P. malariae*	*P. malariae*-like
Assembly							
-Size	23.3	24.4	29.1	33.5	33.5	33.6	23.7
-Scaffolds[a]	14(0)	14(297)	14(226)	14(638)	14(771)	14(47)	14(36)
-Gaps	0	98	560	894	1,264	0	3,697
-GC content	0.19	0.39	0.40	0.29	0.29	0.24	0.30
-Isochore	No	No	Yes	Yes	Yes	Yes	N/A
Genes							
-Number	5,355	5,284	6,671	7,198	7,052*	6,591	4,764*
-Pseudogene	153	7	147	494	N/A	628	N/A
-Density	1/4.3kb	1/4.6kb	1/4.5kb	1/5.0kb	1/5.2kb	1/5.6kb	1/5.3kb
-Intron Size (mean)	167bp	275bp	173bp	178bp	N/A	229bp	N/A
Subtelomeric Genes							
-pir	227	67	1,217	1,949	1,375	255	4
-var	103	0	0	0	0	0	0
-SICAvar	0	241	0	0	0	0	0
-STP1	0	0	9	70	94	166	2
-pv-fam-a	3	13	37	41	33	42	7
-pv-fam-c	0	0	30	8	6	3	3
-ETRAMP	13	9	9	7	11	7	4
-PHIST	77	2	27	24	21	10	3
-fam-l	0	0	0	0	0	396	0
-fam-m	0	0	0	0	0	283	1

[a] Unassigned contigs indicated in parentheses
* Non-curated gene-models

was produced using Illumina sequencing from parasites isolated from a chimpanzee co-infected with *P. reichenowi*. The *P. malariae*-like genome was more fragmented than the other assemblies and its 23.7Mb sequence misses most subtelomeric regions. This lack of subtelomeres is likely due to the whole genome amplification step that was performed prior to sequencing,which preferentially amplifies core regions of the genome where the PCR primers are most likely to bind due to the more moderate GC content.

Most of the *P. malariae* genome is collinear with *P. vivax*, however I did see two instances of large reciprocal translocation breakpoints. The chromosomes syntenic to the *P. vivax* chromosomes 6 and 10 have recombined (figure 1.2 a) and a large pericentric inversion has occurred on chromosome 5 (figure 1.2 b). This was confirmed by mapping sequence data from additional *P. malariae* samples

back to the reference assembly and then looking for sequencing reads that span the recombination breakpoints . Across the four genomes, between 4,764 and 7,198 genes were identified using a combination of *ab initio* gene prediction and projection of genes from existing *Plasmodium* genome sequences. Manual curation performed by Ulrike Boehme was used to correct 2,516 and 2,424 genes for both the *P. malariae* and *P. o. curtisi* reference genomes respectively.

1.5 Comparison to Alternatives

Concurrently to the analysis in this chapter having been performed, draft genomes for both *P. ovale* and *P. malariae* were published by another research group [10]. In comparison to the draft genomes produced by Ansari et al. [10], three of the genomes assembled here (PmUG01, PocGH01, PowCR01) are similar in size but significantly more contiguous (table 1.3). The genome of a chimpanzee-infecting species known as *P. malariae*-like (PmlGA01) is unique to the present study but lacks good coverage of the subtelomeric regions due to biased template representation introduced by the whole genome amplification process. The assembly of *P. malariae* PmUG01 is based on long reads and comprises just 63 pieces. It has no gaps and surpasses the other assemblies according to all metrics reported. The assemblies of the present study are more contiguous, especially in the subtelomeric regions of the genomes.

The manual curation of gene models in the present study made a clear difference to the annotation. In addition to the annotation of pseudogenes, it enabled the identification of approximately 10% more genes as clear 1:1 orthologues of genes in both *P. falciparum* and *P. vivax*. Indeed, in terms of this metric, other available assemblies [10] are similar to the draft assembly of *P. malariae*-like. The highly conserved genes are especially important for cross-species comparisons and analyses. Using 1:1

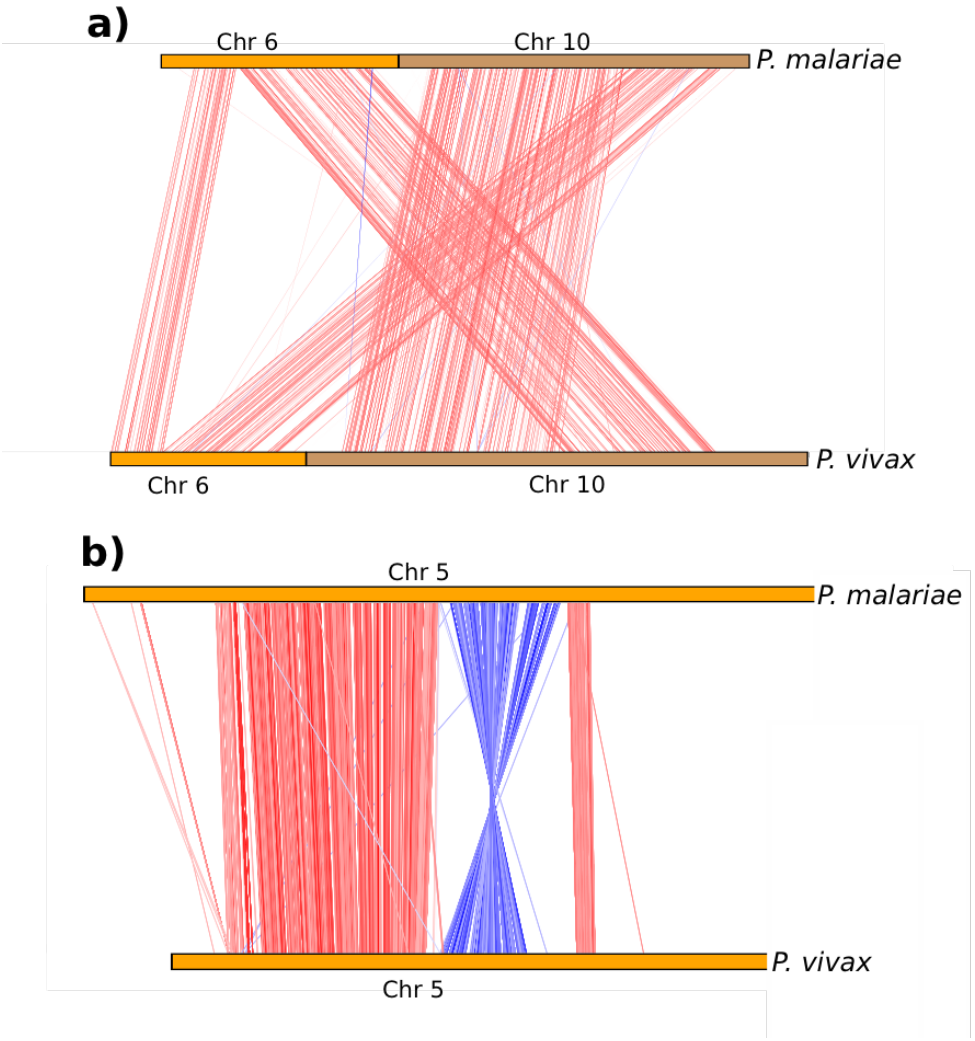


Figure 1.2: Reciprocal translocation breakpoints in P. malariae compared to P. vivax. a) ACT[55] view showing recombination of chromosomes 6 and 10 in P. malariae. The red lines indicate blast similarities, chromosome 6 in orange and chromosome 10 in brown. b) ACT[55] view showing a pericentric inversion in chromosome 5 of P. malariae. Red lines indicate blast similarities and blue lines indicate inverted blast hits.

Table 1.3: Assembly and annotation statistics for the recently described assemblies compared to the present assemblies

	PmUG01	Pmal	PmlGA01	PocGH01	Poc1	Poc2	PowCR01	Pow1	Pow2
Size (kb)	33,618	31,925	23,693	33,485	34,519	38,010	33,579	35,285	35,192
Largest (kb)	3,564	56	3,177	2,946	94	491	3,061	569	657
Average (kb)	534	4	474	22	9	17	43	26	22
Gaps	0	2,236	3,697	894	1,224	2,049	1,264	62	79
Scaffolds	63	7,270	50	654	4,025	2,227	787	1,362	1,611
Scaffolds N50 (kb)	2,312	6	2,076	1,039	18	46	990	174	137
Contigs	63	9,506	3,717	1,548	5,249	4,276	2,047	1,424	1,687
Contigs N50 (kb)	2,312	5	14	39	12	17	30	140	114
Genes	6,591	6,343	4,764	7,198	7,776	8,625	7,052	8,421	8,646
1:1 Orthologs	4,291	3,783	3,837	4,296	3,956	3,874	4,174	3,950	3,958
Core**									
Short Genes[a]	102	104	109	99	69	63	88	89	85
Partial	2	551	90	18	252	201	7	4	4
Pseudo	20	0	245	10	0	0	322	0	0
Unknown function[b]	1,753	1,866	1,508	1,761	1,833	1,804	1,562	1,780	1,778
>7 exon orthologs	281	204	190	280	241	251	260	252	253
Median length (>7 exon) (aa)	477	368	340	478	500	495	462	455	443
Subtelomeres**									
Short Genes[a]	46	278	117	71	536	531	131	857	997
Partial	8	621	246	262	547	676	156	2	6
Pseudogenes	1,236	3	21	978	4	6	393	11	10
Unknown function[b]	765	1,328	447	437	1,176	1,330	734	1,824	2,122

** Core defined as genes that have 1–1 orthologues between *P. falciparum* 3D7 and *P. vivax* P01.

[a] Less than 100 amino acids.

[b] Annotated as either 'hypothetical protein' or 'conserved Plasmodium protein'.

orthologues as indicators of genes within the conserved core regions of chromosomes, I see that my assemblies have about 20% more short genes (less than 100 codons) annotated (averages: 100 versus 82), being thereby more similar in number to those seen in *P. falciparum* (101). Multi-exon genes are notoriously difficult to annotate; looking at the number of 1:1 orthologues in the different assemblies to *P. vivax* and *P. falciparum* genes with over 7 exons (300), the genomes presented here (excluding PmlGA01) have both 10% more 1:1 orthologues annotated and less variable median lengths between assemblies (range: 462-478 versus 368-500[10]). The large number of partial genes observed in some of the genomes assembled by Ansari et al.[10] is due to the higher amount of genes truncated by contig boundaries.

The subtelomeres in *P. malariae* and *P. ovale* required significant manual curation due to the high number of pseudogenes present and the ease in which exons can mistakenly be missed during annotation. Excluding PmlGA01 that lacks most subtelomeric regions, all assemblies presented here have significantly more genes annotated as pseudogenes than those by Ansari et al.[10] (averages: 869 versus 7). The latter study also reports more short genes than my assemblies (averages: 640 versus 81), suggesting potential problems with gene models. Finally, the high gene numbers reported for the assemblies in Ansari et al.[10] can largely be attributed to putative subtelomeric genes, most of which are short with no assigned function and therefore have an increased likelihood of being spurious.

1.6 Phylogenetics

A maximum likelihood tree was constructed using 1,000 conserved core genes that are present as single copies in 12 selected *Plasmodium* species (figure 1.1 b). The four newly assembled genomes did not cluster with any other *Plasmodium* species, but formed two distinct and novel clades. Similar to

a recent study using apicoplast data[14], the two *P. ovale* species formed a sister clade with the rodent malaria species, the latter being an ingroup to the 'superfamily' of primate-infective species in this tree. A further exploration of the phylogenetic tree I report here is detailed in Appendix B. I also saw that *P. malariae*-like has a longer branch length than *P. malariae*, which may be a reflection of the higher levels of diversity in *P. malariae*-like (figure 1.3 a). This lack of diversity in *P. malariae* compared to the chimpanzee species mirrors the situation of *P. falciparum* with *P. reichenowi*[338], another chimpanzee-infective species.

I estimated the time of divergence for the four species using a Bayesian inference tool, G-PhoCS[125]. G-PhoCS uses a large number of unlinked neutrally evolving loci and a given phylogeny to estimate demographic parameters based on the coalescent theory[125]. Absolute divergence time estimates are inherently uncertain due to mutation rate and generation time assumptions, and I therefore scaled these parameters to date the *P. falciparum* and *P. reichenowi* split using G-PhoCS to approximately 4 million years ago (MYA), as previously published (3.0 - 5.5MYA)[322]. Assuming that the mutation rates and generation times are similar for *P. ovale* and *P. falciparum*, I find that the relative split of the two *P. ovale* species is about 5-times earlier than the split of *P. falciparum* and *P. reichenowi*. Using the same mutation rate and generation time as I used to calibrate the *P. falciparum*/*P. reichenowi* split to ≈4MYA, I thereby date the divergence of the two *P. ovale* subspecies to approximately 22.8MYA. Due to being based on non-coding elements, this divergence time is not proportional to the branch lengths of the phylogenomic tree in figure 1.1 b, where coding regions were used. The large divergence time strongly supports the classification of *P. o. curtisi* and *P. o. wallikeri* as separate species rather than subspecies of *P. ovale*.

Using the same mutation rate and a longer generation time to account for the longer intra-erythrocytic

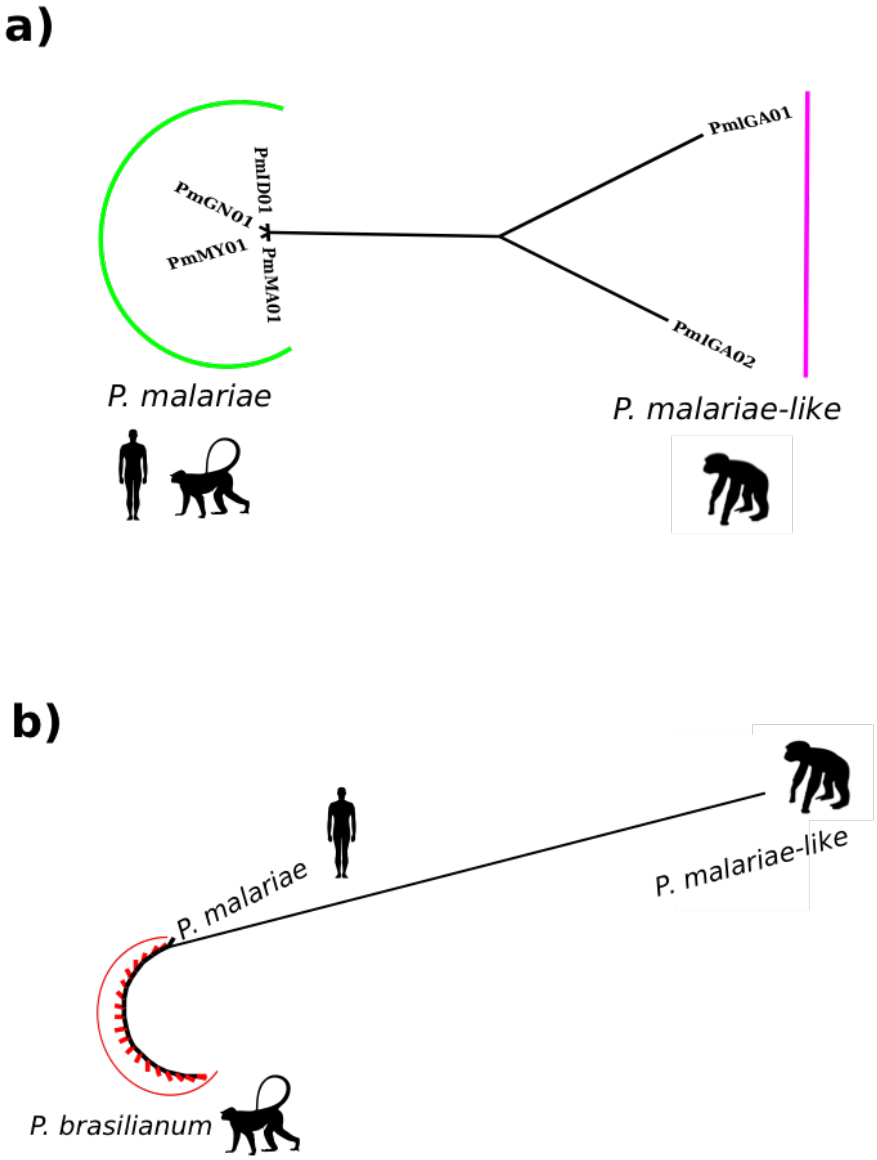


Figure 1.3: P. malariae-like has significantly longer branch lengths than P. malariae, and P. brasilianum is identical to P. malariae. a) A phylogenetic tree of all P. malariae and P. malariae-like samples generated using PhyML [128] based on all P. malariae genes. P. malariae samples are indicated by a green bar and P. malariae-like samples are indicated by a purple bar. Silhouettes represent infectivity. b) A PhyML [128] phylogenetic tree of all P. brasilianum 18S rRNA sequences [177], indicated by a red bar and red tree tips, and the corresponding 18S rRNA sequences from the P. malariae and P. malariae-like assemblies, labeled as such. Silhouettes represent the host origin for each sample.

cycle, I dated the split of *P. malariae* from *P. malariae*-like to ≈3.9MYA. This is similar to the estimated divergence of *P. falciparum* and *P. reichenowi*, suggesting a significant evolutionary event that promoted speciation in *Plasmodium* at that time. It has been suggested that a New World primate-infective species termed *P. brasilianum* is the same species as *P. malariae*[177]. To investigate this further using the new genome assemblies, I aligned the *P. brasilianum* ribosomal rRNA[177] genes to both the *P. malariae* and *P. malariae*-like orthologous genes, showing that the *P. brasilianum* genes are identical to those of *P. malariae*, but that *P. malariae*-like is indeed very different (figure 1.3 b). Subsequent to this analysis, using the recently published *P. brasilianum* genome sequence[344], I found that over 16Mb of the *P. brasilianum* genome sequence matches that of *P. malariae* with over 99% nucleotide similarity, further indicating that they are likely to be the same species.

1.7 Gene Changes

The greater number of genes in both *P. malariae* and *P. ovale* compared to existing *Plasmodium* genomes is mostly due to gene family expansions in the subtelomeres, such as *Plasmodium* interspersed repeat (*pir*) and *STP1* genes (table 1.2). In addition, on chromosome 14, a large expansion was identified in *P. malariae* that comprised 22 tandemly duplicated genes (including two pseudogenes), orthologous to a single *P. falciparum* gene encoding gamete antigen 25/27 (*pfg27*) (figure 1.4 a). *P. vivax* and *P. falciparum* only have one and two copies respectively. *pfg27* is expressed highly during early gametocytogenesis[201], and is essential for correct gametocyte development[252]. This gametocyte gene expansion may be an adaption by *P. malariae* to ensure sexual reproduction in low parasitaemia infections.

In the *P. ovale* species, certain genes are also tandemly duplicated. Nine homologs (including two

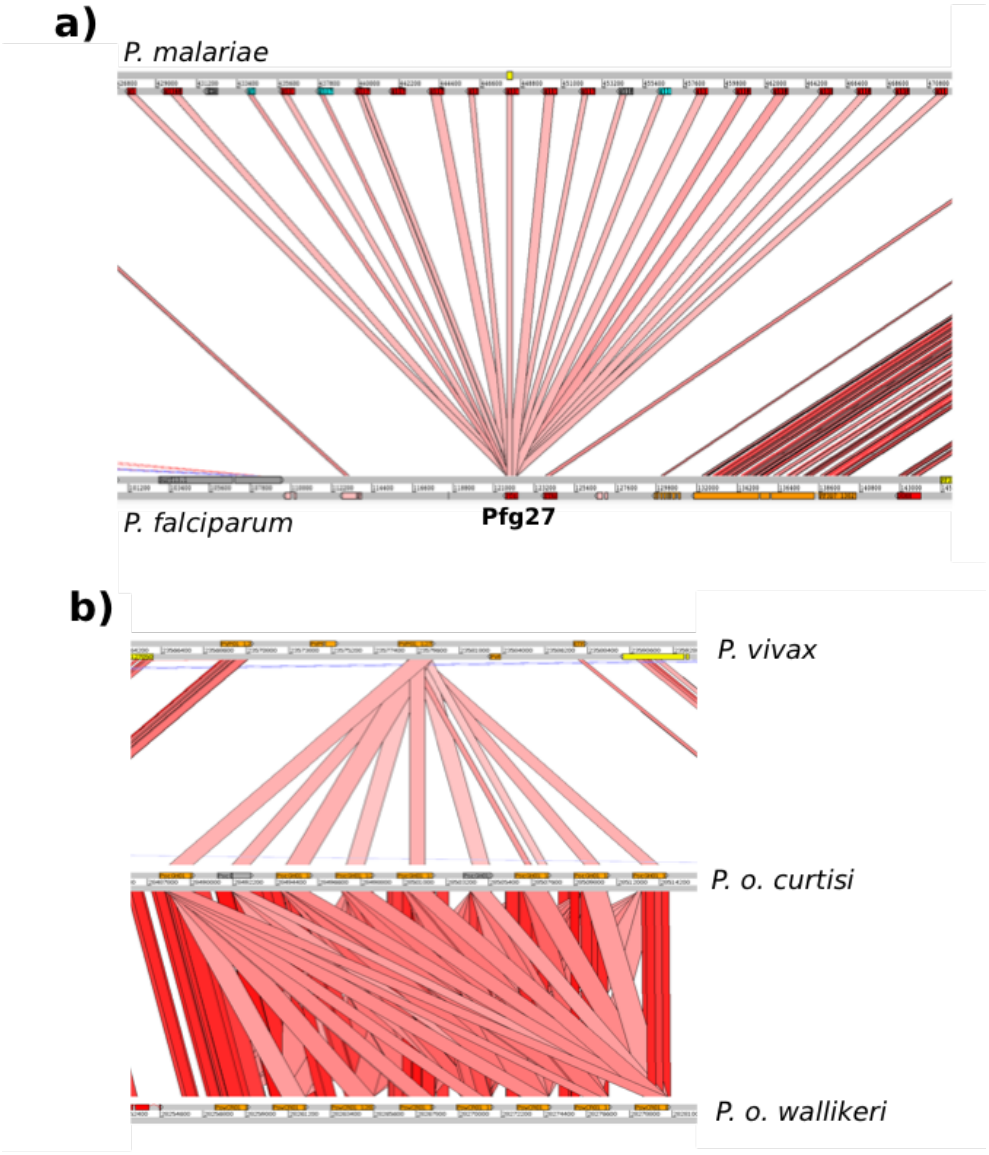


Figure 1.4: Large gene duplications in P. malariae and P. ovale. a) Expansion of pfg27 in P. malariae (top) compared to P. falciparum (bottom) with red lines indicating blast similarities. b) Expansion of PVP01_1270800 (PF3D7_1475900 in P. falciparum), a gene with no known function, in P. o. curtisi and P. o. wallikeri, with different copy numbers in each, compared to the one copy in P. vivax. Functional genes are shown in red and orange (depending on presence/absence of functional annotation) and pseudogenes are shown in grey.

pseudogenes) of PVP01_1270800 are present in *P. o. curtisi* and 7 homologs are present in *P. o. wallikeri* (figure 1.4 b). The *P. vivax* homolog is most highly expressed in sporozoites but has no known function[203]. Using I-TASSER[380], the predicted 3D structure of this gene appears to be similar to a nuclear pore complex (TM-Score > 0.4), suggesting a role in transport. This potential sporozoite change may be indicative of differences in liver-stage invasion or possibly hypnozoite formation.

Multiple genes have become pseudogenes in the two reference genomes compared to other human-infective *Plasmodium* species (table 1.4), *ie.* no canonical structure could be predicted for the genes due to in-frame stop codons, frame shifts or a truncated coding sequence. These pseudogenized genes include homologs of a multidrug efflux pump gene (PF3D7_0212800) which suggests that these species might have a higher innate susceptibility to some drugs. A phosphofructokinase, central to glycolysis, appears to be a pseudogene in both *P. ovale* species, suggesting an alternate route (*e.g.* the pentose phosphate pathway) for glucose turnover in these species. I also saw orthologous genes that are pseudogenes in *P. o. wallikeri* but not in *P. o. curtisi*, such as a serine-threonine protein kinase and a reticulocyte binding protein 1b (RBP1b), that is also pseudogenized in *P. malariae* as discussed below. One gene of specific interest that is pseudogenized in *P. o. wallikeri* but not in *P. o. curtisi* is a homolog of a cyclin in *P. falciparum* (PF3D7_1227500), an observation that may explain the different relapse times of the two *P. ovale* species[244]. The highest number of pseudogenes is seen in the *P. malariae* subtelomeres, where 40% of the genes are pseudogenized in this species, indicating reduced selection pressure to cleanse the genome of these remnant genes, or the pseudogenes potentially being a reservoir for recombination and diversification.

P. malariae has a significantly longer intra-erythrocytic lifecycle compared to other human-infective *Plasmodium* species. All three *Plasmodium* cyclins[295] are highly conserved in *P. malariae*, suggesting

Table 1.4: Pseudogenized and deleted core genes in the two reference genomes

P. vivax ID	Annotation	*P. malariae*	*P. ovale curtisi*
PVP01_0412100	Multidrug efflux pump	Pseudo	Pseudo
PVP01_0309300	Erythrocyte vesicle protein 1	Pseudo	
PVP01_1032500	Conserved Plasmodium protein, unknown function	Pseudo	
PVP01_1344900	Serine/Threonine protein phosphatase CPPED1	Pseudo	
PVP01_1407400	MORN repeat protein	Pseudo	
PVP01_1107900	6-cysteine protein (P92)	Deleted	
PVP01_1117100	Conserved Plasmodium protein, unknown function	Pseudo	
PVP01_0906000	WD repeat-containing protein WRAP73	Deleted	
PVP01_0929100	6-phosphofructokinase		Pseudo
PVP01_0940700	Carbonic anhydrase	Deleted	Pseudo
PVP01_1445600	Conserved Plasmodium protein, unknown function		Pseudo
PVP01_1237400	Nucleoside Transporter 3		Pseudo
PVP01_1123700	Conserved Plasmodium protein, unknown function	Pseudo	Pseudo
PVP01_1246900	Biotin protein ligase		Pseudo

The first column shows the gene identifier of the *P. vivax* P01 homolog of the gene pseudogenized/deleted in one or more of the two reference genome assemblies. The second column is the *P. vivax* P01 annotation of that gene. The following two columns show whether the gene is functional (blank), pseudogenized (Pseudo) or deleted (Deleted).

that the genetic cause may be elsewhere. A WD repeat-containing protein (WRAP73) is deleted in *P. malariae* but conserved across all other *Plasmodium* species. It is part of a large gene family known to be involved in a number of cellular processes, including cell cycle progression[64].

Both *P. ovale* species are able to form hypnozoites, similar to *P. vivax*[51] and the simian-infective *P. cynomolgi*[341]. Of 64 genes exclusive to these hypnozoite-forming species, two genes (table 1.5) do not belong to subtelomeric gene families, encode proteins with transmembrane domains and have orthologs expressed in *P. vivax* sporozoites[366]. The product of one of the two genes has weak similarity to the *P. falciparum* ring-exported protein 4. Looking at genes previously suggested to be involved in hypnozoite formation[341], I did not find *P. ovale* orthologs of the three genes shared exclusively by *P. vivax* and *P. cynomolgi* that contain sporozoite-specific ApiAP2 motifs. However, I did find such motifs in six out of nine dormancy related genes identified[341] (table 1.6), including Ran (PocGH01_09023900) previously identified in a *P. vivax* screen for potential hypnozoite genes[51].

Table 1.5: Potential hypnozoite genes in P. ovale curtisi

PVP01 Annotation	*P. vivax*	*P. cynomolgi*	*P. ovale curtisi*
Ring-exported protein 4*	PVP01_0623900	Pcyb_063280	PocGH01_00129400
Conserved Plasmodium protein	PVP01_1402600	Pcyb_141110	PocGH01_00080600

*not in the same orthologous group as *P. falciparum* REX4.
These are the two orthoMCL gene clusters that contain exclusively all hypnozoite-forming *Plasmodium* species and are not part of subtelomeric gene families.

Table 1.6: Additional hypnozoite gene candidates

P. vivax PVP01	*P. o. curtisi*	Annotation
PVP01_0726200	PocGH01_07035100	Serine/threonine protein phosphatase 4
PVP01_0825700	PocGH01_08034100	Serine/threonine protein phosphatase 6
PVP01_0918300	PocGH01_09023900	GTP-binding nuclear protein Ran/TC4
PVP01_1115000	PocGH01_00015700	Protein kinase 5
PVP01_1205500	PocGH01_12013900	Tyrosine kinase-like protein
PVP01_1257700	PocGH01_12063800	Transcription factor IIIb subunit

Dormancy-related genes identified as containing a sporozoite-specific ApiAP2 motif their 1kb 5' upstream region in *P. vivax*, *P. cynomolgi*, and *P. o. curtisi*.

1.8 Subtelomeric Gene Families

The *Plasmodium* genus is characterized by species-specific subtelomeric gene family expansions, such as *var* genes in *P. falciparum*[336] and *pir* genes in *P. yoelii*[256]. In *P. malariae* and *P. ovale*, where approximately 40% of the total genome size is subtelomeric, large expansions of gene families that are species-specific are also seen (figure 1.5 a) (table 1.2). The three largest gene clusters that I identified were in *P. malariae*. Of these, one cluster is composed of *STP1* genes. Some of these are remarkably similar to surface interspersed genes (*surfins*) in *P. falciparum*[371], further supporting the proposed close relationship of these two gene families[108].

The other two large *P. malariae* clusters consist of two novel gene families, here termed *fam-l* and *fam-m*, consisting of 346 and 283 two-exon ≈250 amino acid long genes respectively. Despite the assembly lacking the majority of the subtelomeres, I also found a *fam-m* gene in the genome of *P.*

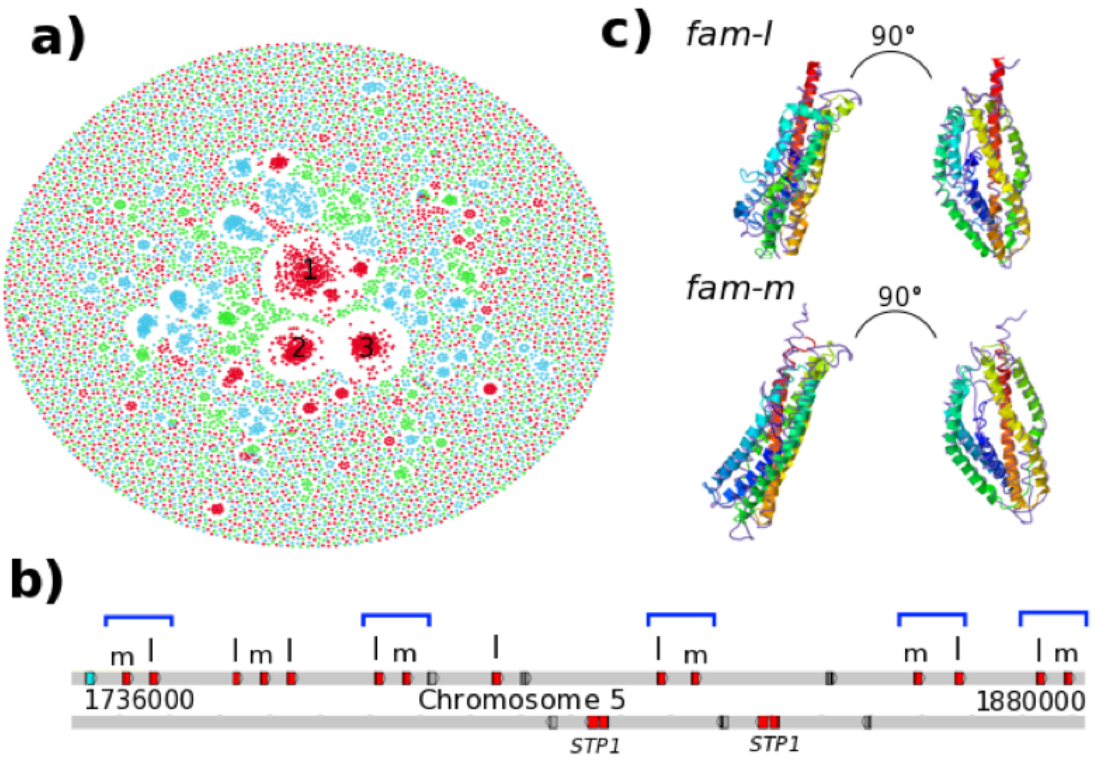


Figure 1.5: Subtelomeric gene family expansions in P. malariae and P. ovale. a) Gene network based on sequence similarity of all genes in P. malariae (Red), P. ovale (Blue), and P. vivax (Green). Cluster 1 contains fam-l genes, Cluster 2 contains fam-m genes, and Cluster 3 contains surfin and STP1 genes. b) Chromosome 5 subtelomeric localization of fam-l and fam-m genes in doublets (Blue brackets) on the telomere-facing strand. Also showing pseudogenes (Grey) and hypothetical gene (Blue). c) Predicted 3D-structure of fam-l (above) and fam-m (below) overlaid with the RH5 crystal structure (Purple). Left images show front of protein, right images show protein tilted to the right.

malariae-like, suggesting that this species also contains at least one member of these novel families. The first exon of each *fam-l* and *fam-m* gene contains a signal peptide and a PEXEL motif –the signature in *P. falciparum* for export from the parasite into host erythrocytes[139]. In addition, the second exon contains two transmembrane domains flanking a hypervariable region. The remainder of the gene sequence is conserved between members of the same family and differentiates the two families from each other. These characteristics support the notion that the proteins encoded by these genes are exported from the parasite and may be targeted to the infected red blood cell surface and play a role in host-parasite interactions.

Ninety-three percent of *fam-l* and *fam-m* genes are on the same strand facing the telomeres (figure 1.5 b). This pattern, similar to *pir* genes in *P. yoelii*[256], may be an adaptation to facilitate recombination between these genes. Uniquely, 60% of these new genes are found as *fam-l* and *fam-m* doublets (figure 1.5 b). Mirror tree analysis suggests that the pairs may be co-evolving over short periods of time (figure 1.6 a), likely through being duplicated together, but that pairing may be disrupted by recombination over longer periods. I did not see any evidence of co-evolution between *pir* genes in close proximity of *fam-m* genes (figure 1.6 b), supporting the fact that this is not an artifact from their subtelomeric location. This suggests that proteins encoded by *fam-l* and *fam-m* genes may form heterodimers when they are exported, a feature not previously seen among subtelomeric gene families in *Plasmodium*.

Finally, I used I-TASSER[380] to predict the 3D structure from both a *fam-l* and a *fam-m* encoded protein. High-confidence (TM score > 0.5) structures were predicted in both cases. These structures overlap the crystal structure of the *P. falciparum* RH5 protein well (TM score > 0.8), with 100% of the RH5 structure covered even though they only have 10% sequence similarity (figure 1.5 c). RH5 is a

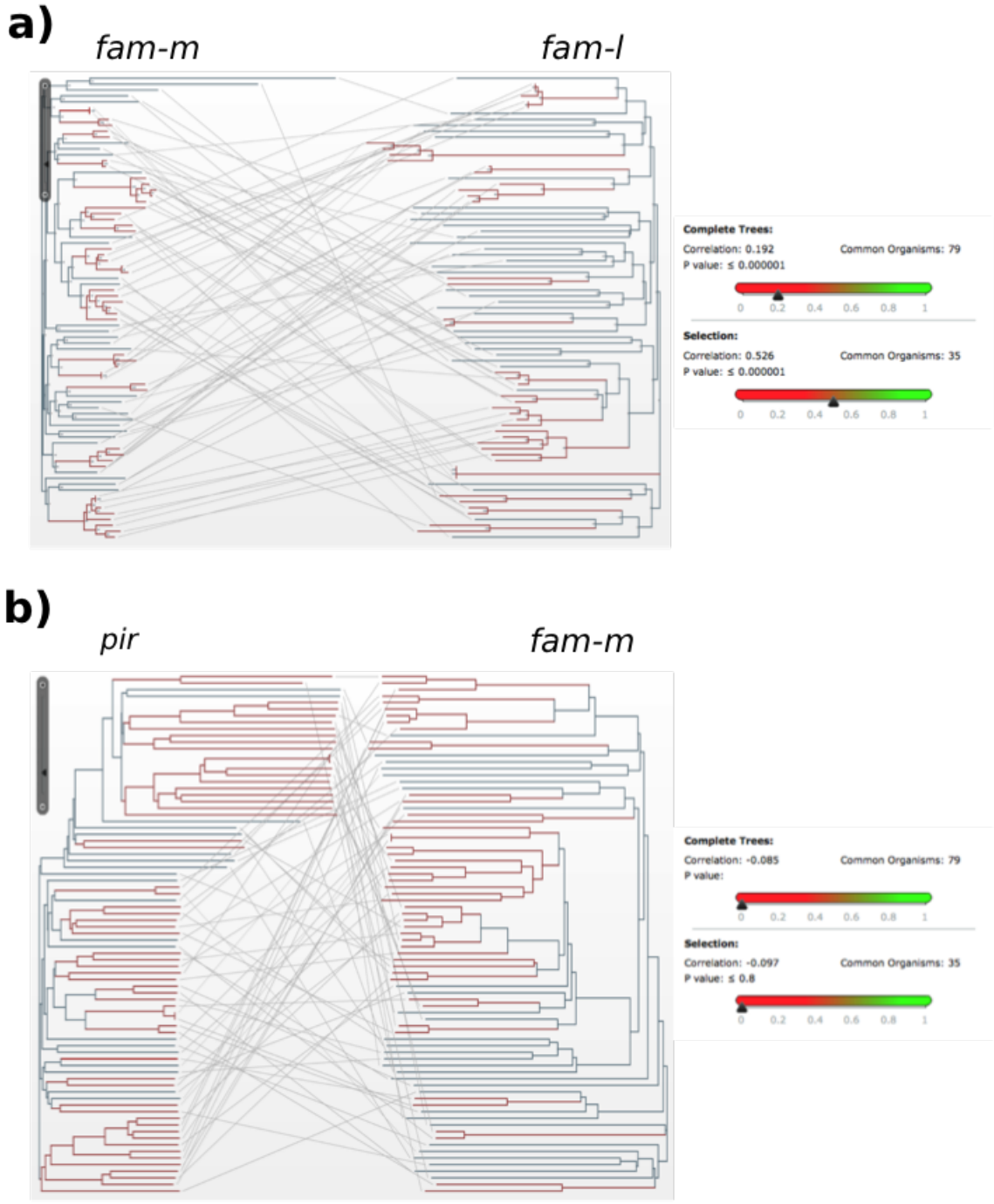


Figure 1.6: Co-evolution of fam-m and fam-l genes, but not with pir genes. a) Mirror tree [248] for 79 fam-m and fam-l doublets, where the two phylogenetic trees correspond to either of the families with lines connecting branch tips of the same doublet. 35 branches (Red) were manually selected due to exhibiting recent branching. Inset shows the correlation between the two trees for all branches (above, r^2=0.19, $p < 0.001$) and red branches (below, r^2=0.53, $p < 0.001$). This shows that the two families are co-evolving, especially when doublets that recently branched are selected. b) Mirror tree [248] for 79 pir and fam-m pseudo-doublets (Appendix A), where the two phylogenetic trees correspond to either of the families with lines connecting branch tips of the same doublet. 35 branches (Red) were manually selected due to exhibiting recent branching. Inset shows the correlation between the two trees for all branches (above, r^2=0.09, $p > 0.05$) and red branches (below, r^2=0.10, $p > 0.05$). This shows that the two families are not co-evolving, and that subtelomeric location does not produce sporadic signals of co-evolution.

prime vaccine target in *P. falciparum* due to its essential binding to human basigin during invasion[74]. The RH5 kite-shaped fold is known to be present in RBP2a in *P. vivax*[127], and may be a conserved structure necessary for the binding capabilities of all RH and RBP genes. This suggests that *fam-l* and *fam-m* encoded proteins also have an adhesion role, possibly binding host receptors.

While neither *P. ovale* species has *fam-l* or *fam-m* genes, they both have large expansions of the *pir* gene family with 1,949 and 1,375 *pir* genes in *P. o. curtisi* and *P. o. wallikeri* respectively, while *P. malariae* only has 255 *pir* genes. This is the largest number of *pir* genes in any sequenced *Plasmodium* genome to date, explaining the large subtelomeres of this species. These *pir* genes form large species-specific clusters suggesting recent expansions (figure 1.5 a), but most closely resemble those in *P. vivax*. An analysis performed together with Adam Reid from the Wellcome Sanger Institute found that many subfamilies of *pir* genes in *P. malariae* and *P. ovale* are shared with *P. vivax*, while almost none are shared with the rodent-infecting species (figure 1.7 a). This suggests that *pir* genes are relatively well conserved between non-*falciparum* species infecting humans. Interestingly, all hypnozoite-forming species (both *P. ovale*, *P. vivax*, and *P. cynomolgi*) contain over 1,000 *pirs* each, significantly more than non-hypnozoite-forming *Plasmodium* species. Using additional draft genome assemblies for both *P. o. curtisi* and *P. o. wallikeri*, I show that the two species of *P. ovale* share significantly fewer *pir* genes inter-specifically than they do intra-specifically or intra-genomically (99% identity over 150 amino acids), further suggesting that the two species are not recombining with each other (figure 1.7 b).

1.9 Reticulocyte and Duffy Binding Proteins

RBP genes encode a merozoite surface protein family present across all *Plasmodium* species and known to be involved in red blood cell invasion and host specificity[155]. Compared to *P. vivax*, *P. malariae*

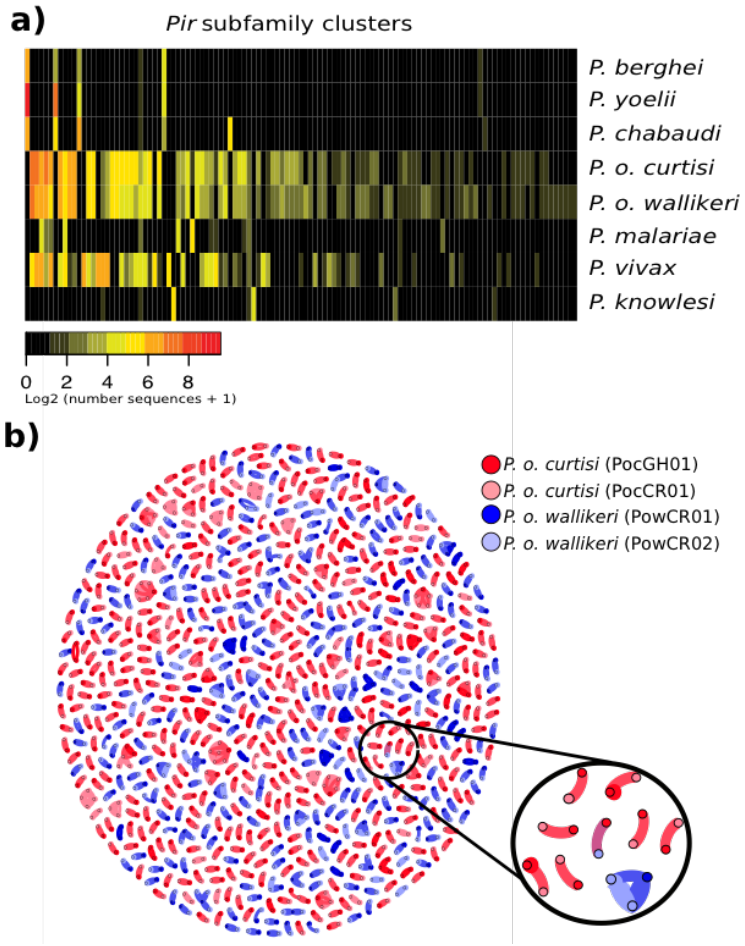


Figure 1.7: pir genes in P. malariae and P. ovale resemble those in P. vivax, and pir genes are less similar between the two P. ovale than within. a) Heatmap showing the sharing of pir subfamilies between different species based on tribeMCL [96]. Columns show pir subfamilies and rows show species. Colours indicate the number of genes classified into each subfamily for each species. Subfamilies were ordered by size, species were ordered for clarity. pir genes in rodent-infecting species fall into a small number of well-defined families. Those in P. vivax, P. malariae and P. ovale are however much more diverse. There is little overlap between rodent subfamilies and human-infecting subfamilies, despite P. ovale being a sister taxa to the rodent-infecting species. P. knowlesi has some sharing with other species, but its largest families are species-specific, suggesting it has undergone specialization of its pir repertoire. b) Gene network of pir genes for both high-quality assemblies of P. o. curtisi (Dark red) and P. o. wallikeri (Dark blue) and draft assemblies of each (Light red and light blue respectively). pir genes with BLASTP [5] identity hits of 99%+ over 150 amino acids become connected in the graph. Genes without connections were removed. There is one connection between the two species (circled in black and with a zoomed in version), 801 between the P. o. curtisi assemblies, 524 between the P. o. wallikeri assemblies, 527 on average within each P. o. curtisi assembly, and 423 on average within each P. o. wallikeri assembly.

has lost multiple RBPs including nearly all RBP2 genes and RBP1b, though it does have a functional RBP3. On the other hand, the two *P. ovale* species each have multiple full-length RBP2 genes (seven in *P. o. curtisi* and four in *P. o. wallikeri*) compared to three copies in *P. vivax* (figure 1.8 a). The two *P. ovale* species have very similar RBP2s, such as PocGH01_00019400 and its ortholog, a number of RBP2 pseudogenes in the two genomes match with a functional copy in the other genome (figure 1.9 a). The RBP1b pseudogene in *P. o. wallikeri* is less degenerate than in *P. malariae* and in *P. malariae*-like where only a short fragment of the gene was found (figure 1.8 b). The specific mutation introducing a stop codon is conserved across the two *P. o. wallikeri* samples (figure 1.9 b), indicating that RBP1b has become pseudogenized recently in this species, or that the shortened form may be functional and has therefore been maintained under selection. It is interesting to note that the positioning of RBP1b and RBP1a is conserved across all these species, but not with the rodent malaria species.

RBP genes are thought to be involved specifically in reticulocyte invasion, which explains the gene loss in *P. malariae*, a species that preferentially invades normocytes[68] (figure 1.8 c). Both *P. ovale* species exclusively invade reticulocytes[67] and may have developed novel invasion pathways through the RBP2 expansion, similar to *P. vivax*. This supports a role for RBP2 gene expansions specifically in reticulocyte invasion. RBP3 genes seem to be pseudogenized in all reticulocyte-infective species, while they are fully functional in normocyte-infective species, suggesting a role in normocyte-invasion for RBP3.

Duffy binding proteins (DBPs) are also important for erythrocyte invasion[155]. *P. malariae* has one functional and one recently pseudogenized DBP, while both *P. ovale* have two functional copies. It is believed that *P. vivax* struggles to infect duffy-negative humans due to relying on its DBP binding

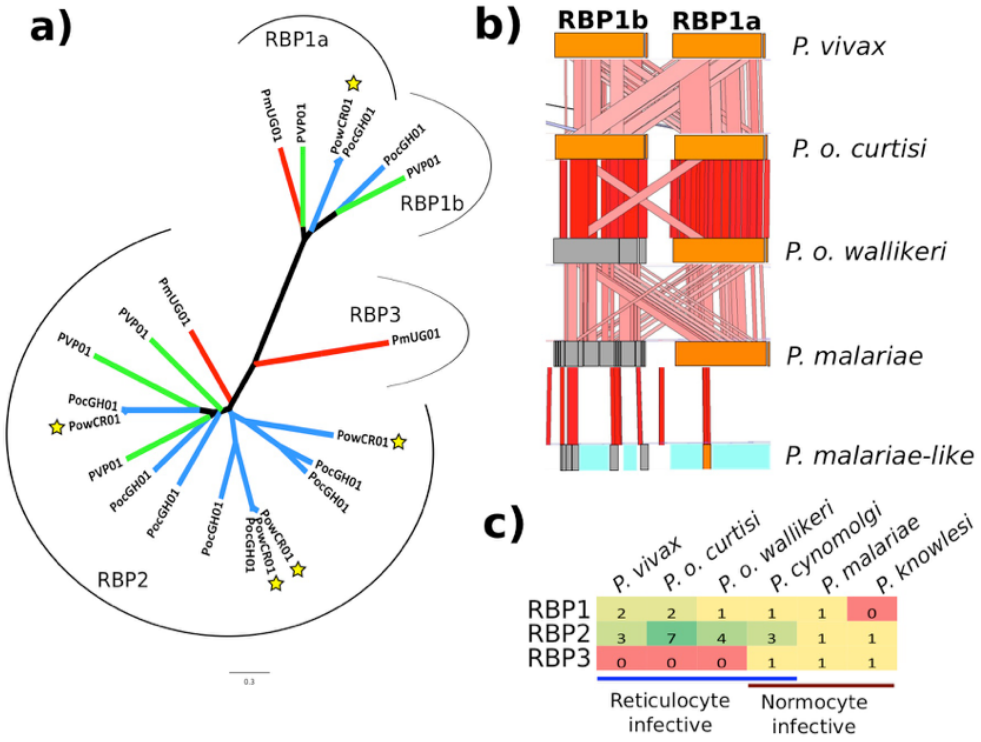


Figure 1.8: Reticulocyte binding protein changes in P. malariae and P. ovale. a) Phylogenetic tree of all full-length functional RBPs in P. malariae (Red branches), P. o. curtisi (Blue branches without stars), P. o. wallikeri (Blue branches with stars), and P. vivax (Green branches). Brackets indicate the different subclasses of RBPs: RBP1a, RBP1b, RBP2, and RBP3. b) ACT[55] view of functional (Orange) and pseudogenized (Grey) RBP1a and RBP1b in five species (P. vivax, P. o. curtisi, P. o. wallikeri, P. malariae, P. malariae-like). Blue indicates assembly gaps. Red bars between species indicate level of sequence similarity, with darker colour indicating higher similarity. c) Number of RBP genes in each of the three RBP classes (RBP1, RBP2, RBP3) by species (P. vivax, P. o. curtisi, P. o. wallikeri, P. cynomolgi, P. malariae, P. knowlesi) grouped by erythrocyte invasion preference (reticulocyte versus normocyte).

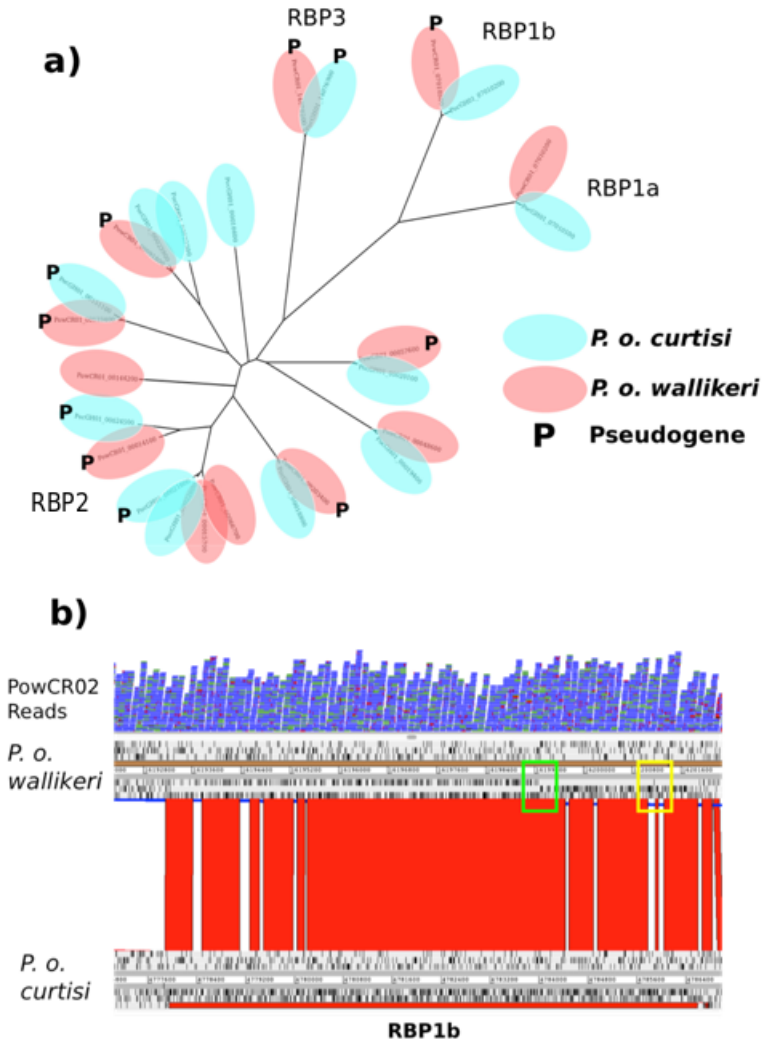


Figure 1.9: Multiple RBP-encoding genes are pseudogenized between the two P. ovale species. a) PhyML [128] generated phylogenetic tree of all RBP-encoding genes over 1kb long in P. o. curtisi (light blue) and P. o. wallikeri (light red). Pseudogenes are denoted with **P**. Multiple functional RBP2 genes match up with pseudogenized copies in the other genome. The gene IDs in the figure are not meant to be legible. b) ACT [55] view of RBP1b in red for P. o. curtisi (bottom) and the corresponding disrupted open reading frame in P. o. wallikeri (top), with black ticks indicating stop codons. Reads (in blue) from an additional P. o. wallikeri sample (PowCR02) confirm the bases introducing the frameshift (green square) and premature stop codon (yellow square) in RBP1b.

the Duffy antigen, with recent studies showing an association between duffy-negative infectivity and *P. vivax* strains containing a DBP duplication[218]. The fact that *P. malariae* and *P. ovale* are found throughout Africa (figure 1.1 a) suggests that they are capable of infecting duffy-negative individuals. This implies that one DBP copy appears to be sufficient for *P. malariae* to infect humans who are duffy-negative.

1.10 Differential Selection Pressures

Using four additional *P. malariae* samples, two additional *P. o. curtisi* samples and two *P. malariae*-like and *P. o. wallikeri* samples each (appendix table A.1), I investigated differences in selection pressures between two species that diverged based on host differences (*P. malariae* and *P. malariae*-like), and two species that supposedly diverged within the same host (*P. o. curtisi* and *P. o. wallikeri*). GATK's UnifiedGenotyper[215] was used to call SNPs and, following standard filtering (Appendix A), I retained 230,881 SNPs in *P. malariae* and 1,462,486 SNPs in *P. ovale* (tables 1.7 & 1.8). Excluding subtelomeric regions, the pairwise nucleotide diversity between the different *P. malariae* samples is 3.2×10^{-4} and for the *P. o. curtisi* samples it is 1.9×10^{-4}, which are lower than the estimates we obtained for *P. vivax* (9.9×10^{-4}) and *P. falciparum* (5.7×10^{-4}) using the same methodology. The nucleotide diversity for *P. malariae*-like is 6.5×10^{-3}. Interestingly, the nucleotide diversity of *P. o. wallikeri* (3.7×10^{-4}) appears to be much higher than that of *P. o. curtisi*, though this is difficult to confirm due to low sample numbers.

Every core gene with more than 5 nucleotide substitutions and which had identifiable orthologs in *P. falciparum* and *P. vivax* (2,343 genes in *P. malariae*, 4,023 genes in *P. o. curtisi*) was analysed for evidence of selection, using the following approaches: the Hudson-Kreitman-Aguade ratio (HKAr)[152],

Table 1.7: P. malariae & P. malariae-like SNP Calling Results

Sample ID	PmMY01	PmID01	PmMA01	PmGN01	PmlGA01[a]	PmlGA02[a]
Raw SNPs	218,334	164,541	173,028	239,655	458,790	211,686
- Private	48,094	19,475	25,901	66,377	260,540	68,793
- Ref	712,758	696,634	706,817	737,236	386,915	261,042
- Missing*	50,394	86,900	74,936	28,776	165,813	415,250
Filtered SNPs	8,970	8,589	7,742	7,878	161,551	140,113
- Private	2,149	2,066	1,908	2,066	77,781	56,571
- Ref	221,923	222,247	223,058	223,003	69,466	90,571
- Missing*	0	0	0	0	0	0

[a] *P. malariae*-like sample
*sites at which the sample has no coverage
SNP calling results as per mapping all *P. malariae* and *P. malariae*-like samples against the PmUG01 PacBio reference genome assembly. The raw SNPs are the total number of SNPs that I called using GATK [215] default parameters in the different samples. Of these raw SNPs, some are exclusive to a certain sample (Private), are identical to the reference genome (Ref), or there is no coverage and therefore no SNP call could be made (Missing). The same information is also shown for the filtered SNPs, which were filtered according to a number of different parameters (Appendix A).

Table 1.8: P. ovale curtisi & P. ovale wallikeri SNP Calling Results

Sample ID	PocGH02	PocCR01	PowCR01	PowCR02
Raw SNPs	171465	277978	2139946	1881088
- Private	36487	99083	333727	83609
- Ref	2287008	2249682	693405	674071
- Missing*	84743	72495	104013	149166
Filtered SNPs	29099	46695	1415164	1410434
- Private	6162	16026	21081	16699
- Ref	1433387	1416042	45978	50845
- Missing*	0	0	0	0

[a] *P. ovale wallikeri* sample
*sites at which the sample has no coverage
SNP calling results as per mapping all *P. o. curtisi* and *P. o. wallikeri* samples against the PocGH01 Illumina reference genome assembly. The raw SNPs are the total number of SNPs that I called using GATK [215] default parameters in the different samples. Of these raw SNPs, some are exclusive to a certain sample (Private), are identical to the reference genome (Ref), or there is no coverage and therefore no SNP call could be made (Missing). The same information is also shown for the filtered SNPs, which were filtered according to a number of different parameters (Appendix A).

which is the ratio of interspecific nucleotide divergence to intraspecific polymorphisms (*ie.* diversifying selection), Ka/Ks[241], to look for an enriched number of nonsynonymous differences compared to synonymous differences (*ie.* positive selection), and the McDonald Kreitman (MK) Skew[174], a measure of maintained polymorphisms (*ie.* balancing selection). Appendix A contains further details on how these selection measures are calculated. Due to differences in denominators for HKAr measures making comparisons between the species pairs invalid, for each species pair I therefore calculated the mean HKAr and used a threshold of two standard deviations above that mean to signify an elevated HKAr. For *P. malariae*/*P. malariae*-like, a threshold of HKA > 0.27 was determined while a threshold of HKA > 0.075 was found for *P. o. curtisi*/*P. o. wallikeri*. Using those thresholds, I found 3.5% of genes (81/2,343) to have an elevated level of HKAr in *P. malariae* but significantly fewer (1.4%; 55/4,023) in *P. o. curtisi* (2-sample test for equality of proportions, $p < 0.001$) (table 1.9). More genes under significant balancing selection were seen in *P. malariae* (17/2,343, 0.7%) than in *P. o. curtisi* (4/4,023, 0.1%) ($p < 0.001$). Additionally, more genes are under strong positive selection (Ka/Ks > 2) in *P. malariae* (131/2,343, 5.6%) than in *P. ovale* (58/4,023, 1.4%) ($p < 0.001$), with most *P. malariae* genes having a higher ratio of nonsynonymous to synonymous fixed mutations compared to *P. o. curtisi* (figure 1.10 a). This genome-wide increase in nonsynonymous fixations is indicative of a population bottleneck, which may underlie the high proportion of genes with signatures of positive or balancing selection in *P. malariae*.

Looking at specific genes under selection, similar genes were identified in the *P. malariae*/*P. malariae*-like test as in an earlier *P. falciparum*/*P. reichenowi* study[259], hinting at conserved selection pressures in speciation between human and chimpanzee hosts (table 1.10). A number of genes have high HKAr values in both comparisons (figure1.10 b), including MSP1 and a number of gameto-

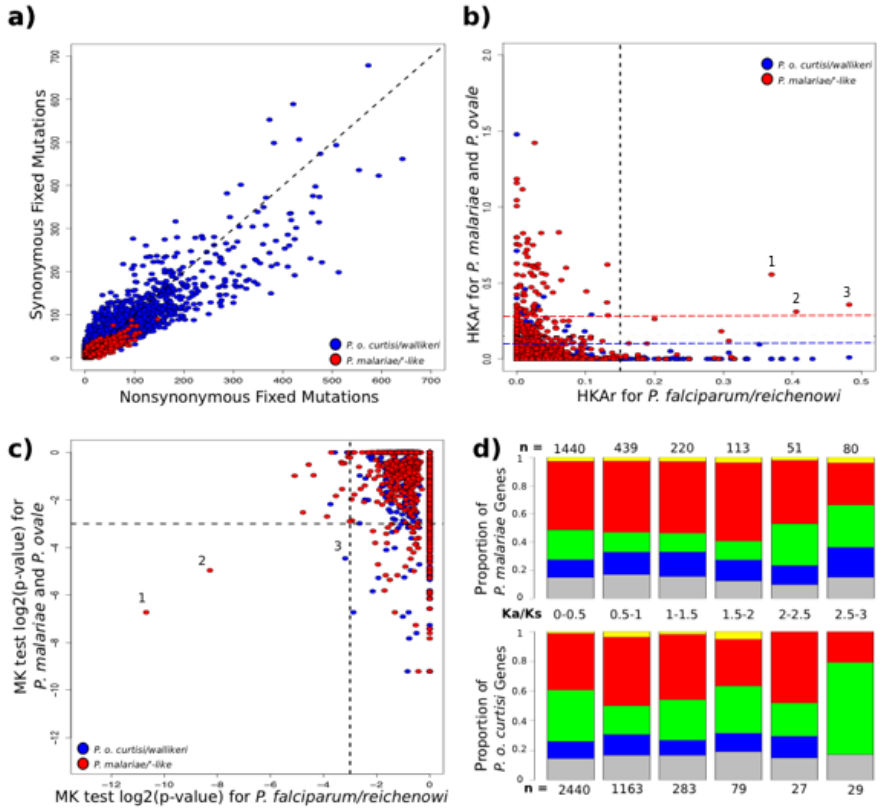


Figure 1.10: Population genetics measures differ between P. malariae and P. o. curtisi. a) Nonsynonymous versus synonymous fixed mutations per gene for both the P. o. curtisi/P. o. wallikeri (blue) and the P. malariae/P. malariae-like (red) comparisons. While the former has most genes centred around the x=y line, the latter has most genes below this line with more nonsynonymous than synonymous mutations, indicative of an ancestral bottleneck. b) Gene-wide HKAr values for the P. falciparum/P. reichenowi comparison, described earlier [259], versus HKAr for the P. o. curtisi/P. o. wallikeri (blue) and the P. malariae/P. malariae-like (red) comparisons (respective thresholds shown in corresponding colours). Three genes show elevated HKAr values for both comparisons: 1) ADP/ATP carrier protein (PF3D7_1004800) 2) merozoite surface protein 1 (PF3D7_0930300) 3) conserved Plasmodium protein (PF3D7_0311000). c) Log_2 of p-values of gene-wide MK tests for the P. falciparum/P. reichenowi comparison [259] by P. o. curtisi/P. o. wallikeri (blue) and P. malariae/P. malariae-like (red) comparisons. Three genes have significant MK skews ($log_2(p) < -3$) for both comparisons: 1) conserved Plasmodium protein (PF3D7_1361800) 2) apical membrane antigen 1 (PF3D7_1133400) 3) NAD(P)H-dependent glutamate synthase (PF3D7_1435300). d) Barplots of proportion of P. malariae (above) and P. o. curtisi (below) genes expressed at different stages (No peak expression (grey), ookinete (blue), gametocyte (green), intraerythrocytic (red), and other stage(yellow)) binned by Ka/Ks, with the number of genes in each bin displayed (n =). P. o. curtisi genes with very high Ka/Ks values (> 2.5) are enriched for gametocyte genes.

Table 1.9: Genes with significant scores in two or more population genetics measures

Species	Gene ID	Gene Product
P. malariae	PmUG01_05040800	hypothetical protein
P. malariae	PmUG01_07023900	alkaline phosphatase
P. malariae	PmUG01_13030900	transcription factor with AP2 domain
P. malariae	PmUG01_14040200	conserved Plasmodium protein
P. malariae	PmUG01_14019500	conserved Plasmodium protein
P. malariae	PmUG01_12012900	conserved Plasmodium protein
P. malariae	PmUG01_07042000	merozoite surface protein 1
P. malariae	PmUG01_10013600	formin 1
P. malariae	PmUG01_13030700	rRNA (adenosine-2'-O-)-methyltransferase
P. malariae	PmUG01_14062900	merozoite surface protein 9

For the three population genetics measures (HKAr, Ka/Ks, and MK Skew), the table shows the genes that have significant values in two or more of these measures. These genes therefore represent genes under significant selection pressures.

cyte/ookinete genes such as EGF-like membrane protein, ferrodoxin reductase-like protein and an ADP/ATP carrier protein. Two blood stage genes have significant MK skews for both comparisons (figure 1.10 c), including a conserved protein of unknown function and apical membrane antigen 1. Amongst genes with significant selection coefficients in both comparisons Gene Ontology (GO) annotations of 'pathogenesis' and 'entry into/exit from host cell' are significantly enriched. Similarly, a number of blood stage genes are found to have both high HKAr and significant MK skews, including MSP1, MSP9, and formin-1, all of which are known to be important in invasion and also have the same GO terms enriched. One of the genes with the highest Ka/Ks in the *P. malariae*/*P. malariae*-like comparison is RBP1a, which has 37 nonsynonymous fixed differences between the two species and only 6 synonymous fixed differences. The other two intact RBPs are much more highly conserved. These data overall suggest that the selection pressures acting on *P. malariae* and *P. falciparum* lineages have been similar since the split from their chimpanzee-infecting relatives, and that this selection is acting primarily on blood stage genes and certain sexual stage genes.

Table 1.10: Genes with significant scores in same test for both P. falciparum/P. reichenowi and either P. o. curtisi/P. o. wallikeri or P. malariae/P. malariae-like

Species	Gene ID	Gene Product
P. malariae	PmUG01_09042600	apical membrane antigen 1
P. malariae	PmUG01_11024300	conserved Plasmodium protein
P. malariae	PmUG01_03026800	ferrodoxin reductase-like protein
P. malariae	PmUG01_07042000	merozoite surface protein 1
P. malariae	PmUG01_08020600	ADP/ATP carrier protein
P. malariae	PmUG01_08045200	conserved Plasmodium protein
P. malariae	PmUG01_11040300	EGF-like membrane protein
P. o. curtisi	PocGH01_13025000	NAD(P)H-dependent glutamate synthase

For the three population genetics measures (HKAr, Ka/Ks, and MK Skew), the table shows the genes that have significant values in both the *P. falciparum/P. reichenowi* comparison and either the *P. o. curtisi/P. o. wallikeri* or the *P. malariae/P. malariae*-like comparison.

As expected, I do not see any significant sharing of selection pressures for the two *P. ovale* species with *P. falciparum/P. reichenowi* (table 1.10), besides an NAD(P)H-dependent glutamate synthase which has a significant MK skew in both comparisons (figure 1.10 c). *P. ovale* genes with significant HKAr values include a number of transporters, including a homolog of an ABC transporter (MRP2), with GO terms enriched for 'drug transmembrane transport' and 'intracellular transport'. However, the five genes with the highest HKAr are all gametocyte and ookinete genes, including among others a transcription initiation factor TFIID and a mago nashi homolog protein, the latter potentially being involved in sex determination [195]. We also find that genes with low Ka/Ks (<0.5) and very high Ks/Ks (>2.5) are enriched for gametocyte genes (hypergeometric test, $p < 0.0001$ and $p < 0.001$ respectively) (figure 1.10 d). Genes with high Ka/Ks values that are gametocyte-associated are enriched for genes of unknown function (hypergeometric test, $p < 0.001$), suggesting important novel *Plasmodium* biology. Of the four genes with significant MK skews in *P. o. curtisi*, one is a kelch protein while the others are involved in 'DNA replication' and 'Telomere maintenance'. These results hint at a number of possible divergences between the two *P. ovale* species, including possible differences in drug sus-

ceptibility, changes in gametocyte genes that may have enabled speciation, while differences in DNA replication may possibly be linked to the different relapse times.

1.11 Conclusion

The high-quality genome sequences of *P. malariae* and *P. ovale* and their annotation presented here provide a rich new resource for comparative *Plasmodium* genomics. They provide a foundation for further studies into the biology of these two neglected malaria species, as well as new tools to explore genus level similarities and differences in infection. The genome sequences have revealed a number of genomic adaptations and possible consequences related to the success of these species in sustaining low parasitaemia infections, including gametocyte gene expansions and an increase in genome size. The genome sequences suggest that the rodent-infective malaria species may be the result of an ancestral host switch from a primate-infective species and also conclusively show that *P. ovale* is a species complex, consisting of two highly diverged species, *P. o. curtisi* and *P. o. wallikeri*. The genome sequences reveal a novel type of subtelomeric gene family in *P. malariae* occurring in doublets and potentially having an RH5-like fold. Having access to a larger number of genome sequences also allows us to identify features such as the RBP2 gene expansion in reticulocyte invading *Plasmodium* species. Multi-sample analysis of the two species highlights differences in selection pressures between host-switching and within-host speciation, as well as the omnipresent selective pressure during red blood cell invasion. These genome sequences will now enable more comprehensive studies of human-infectivity in *Plasmodium* species.

In addition to their important uses in understanding the evolutionary history of the *Plasmodium* genus, these new genome sequences will now also enable us to better understand and characterize the

P. malariae, *P. o. curtisi* and *P. o. wallikeri* infections that are detected in clinical settings. Upon completing the analysis presented in this chapter, I was made aware of the fact that the patient who provided the *P. malariae* sample used for the reference genome assembly had subsequently returned to the hospital with a recrudescent infection of *P. malariae*. Analysing and understanding from a genomic perspective the circumstances that potentially led to this recrudescence is the topic of Chapter 2.

The belief is growing on me that the disease is communicated by the bite of the mosquito. ... She always injects a small quantity of fluid with her bite - what if the parasites get into the system in this manner?

Ronald Ross, Letter to Patrick Manson, 1896 CE

2

A case of clinical treatment failure in a *P. malariae* infection

2.1 Abstract

Plasmodium malariae is the only human malaria parasite species with a 72-hour intraerythrocytic cycle and the ability to persist in the host for life. Here I present a case of a *P. malariae* infection with clinical recrudescence after directly observed administration of artemether/lumefantrine. By using

whole-genome sequencing, I show that the initial infection was polyclonal and the recrudescent isolate was a single clone present at low density in the initial infection. Haplotypic analysis of the clones in the initial infection revealed that they were all closely related and were presumably recombinant progeny originating from the same infective mosquito bite. I review possible explanations for the *P. malariae* treatment failure and conclude that a 3-day artemether/lumefantrine regimen is suboptimal for this species because of its long asexual lifecycle.

2.2 Introduction

During the past decade, intensification of malaria control efforts has substantially reduced the global burden of malaria from *Plasmodium falciparum*. This trend has often been associated with increased recognition of the burden of malarial disease caused by the other *Plasmodium* species[1]. *P. malariae*, one of the six *Plasmodium* species that commonly infect humans, is endemic throughout parts of Africa[222,39], South America[311], Asia, and the Western Pacific[163]. *P. malariae* is unique among the human-infective *Plasmodium* species in having a 72-hour intraerythrocytic lifecycle with variable but often prolonged pre-erythrocytic intrahepatic development[68]. *P. malariae* can persist in the human host for years and possibly an entire lifetime. Although it is often asymptomatic, chronic parasitemia in endemic areas is associated with substantial rates of illness, including anemia and nephrotic syndrome[119,87,178].

A key strategy for malaria elimination is the strengthening of health systems to deliver early diagnosis and highly effective therapy. Artemisinin-based combination therapy (ACT) has been central to this approach, with proven efficacy against multidrug-resistant *P. falciparum*, multidrug-resistant *P. vivax*, and against *P. knowlesi*[246,290,86,124]. In recent years, there have been increasing calls for a uni-

versal policy of ACT for all species of malaria[246,290,86,124]. However, the efficacy of ACT against *P. malariae* is poorly documented.

Although chronic infection with *P. malariae* is well-recognized[19], little is known regarding how the parasites manage to evade host immunity and the intrahost dynamics of the underlying parasite population. In Chapter 1, I described the production and analysis of a new reference genome for *P. malariae*. The *P. malariae* reference genome is 33.6 Mb in size, has 6,540 genes, and has an average GC content of 24%[301].

Here I report a case of a *P. malariae* infection in a patient residing in a non–malaria-endemic environment that resulted in recrudescence months after treatment with artemether/lumefantrine (AL). By using whole-genome sequencing of isolates from the initial and the recrudescent infection, I show that the two major *P. malariae* haplotypes, constituting 90% of the parasite load in the initial infection, were cleared successfully by AL, whereas a third haplotype, constituting a minority subpopulation in the initial infection, survived and recrudesced.

All methods used for this chapter are detailed in Appendix C. The clinical work performed in the 'patient presentation' section was performed by Dr. Ian Marr from the Royal Darwin Hospital in Australia, while the genetic data analyses were all performed by me.

2.3 Patient Presentation

A 31-year-old Uganda-born man, weighing 77 kg (170 lbs), who had been a resident in Australia for 5 years sought care at Royal Darwin Hospital (Darwin, Northern Territory, Australia) on March 1, 2015, with a 4-day history of fevers and headaches. He had returned to Australia 56 days previously after a 2-week trip to Uganda visiting friends and relatives (figure 2.1 a & e). He had spent 14 days in a

rural malaria-endemic area in eastern Uganda. Although he had not taken regular malaria prophylaxis, he had self-medicated with a locally acquired oral course of AL on the second and third day of his trip, despite being clinically well (figure 2.1 b & d). He returned to Australia (now a malaria-free country) in January 2015 until seeking care after a short febrile illness in late February. On examination, he had a tympanic temperature of 37.5°C and a heart rate of 110 beats/min but no manifestations of severe malaria. Rapid diagnostic testing with BinaxNOW (Binax, Inc, Inverness Medical Professional Diagnostics, Scarborough, ME, USA) for malaria was positive for aldolase but negative for histidine-rich protein 2. Species-specific PCR was positive for *P. malariae* and negative for all other *Plasmodium* species. Thick and thin blood film examination confirmed *P. malariae* parasitemia (12,140 parasites/μL) with all stages of asexual development visible on the blood film (figure 2.2 a & b). The blood film was otherwise unremarkable; in particular, no evidence for hyposplenism was found. The patient was not immunosuppressed, and an HIV serologic test was negative. A hepatitis C serologic test was positive but with a viral load that was below the limit of quantification (<12 IU/mL).

The patient was administered a single 20/120 mg tablet of AL on the first day because of a prescribing error but subsequently continued with a supervised standard regimen of 80/480 mg every 12 hours taken with fatty food to complete a full course of 6 doses over 3 days, equivalent to a total dose of 6.2 mg/kg of artemether and 37.4 mg/kg of lumefantrine. Glucose 6-phosphate dehydrogenase function was normal, and a single 30-mg dose of primaquine was administered on day 2. His hemoglobin was 126 g/dL and he received no blood transfusion. After treatment, his parasitemia declined to 1,269/μL at 32 hours, 488/μL at 41 hours, and 55/μL at 56 hours. He was afebrile and symptom-free within 36 hours of admission. However, before discharge on day 6, thick blood film examination was still positive (192/μL), but by day 11, his repeat blood film examination and his aldolase rapid diagnostic

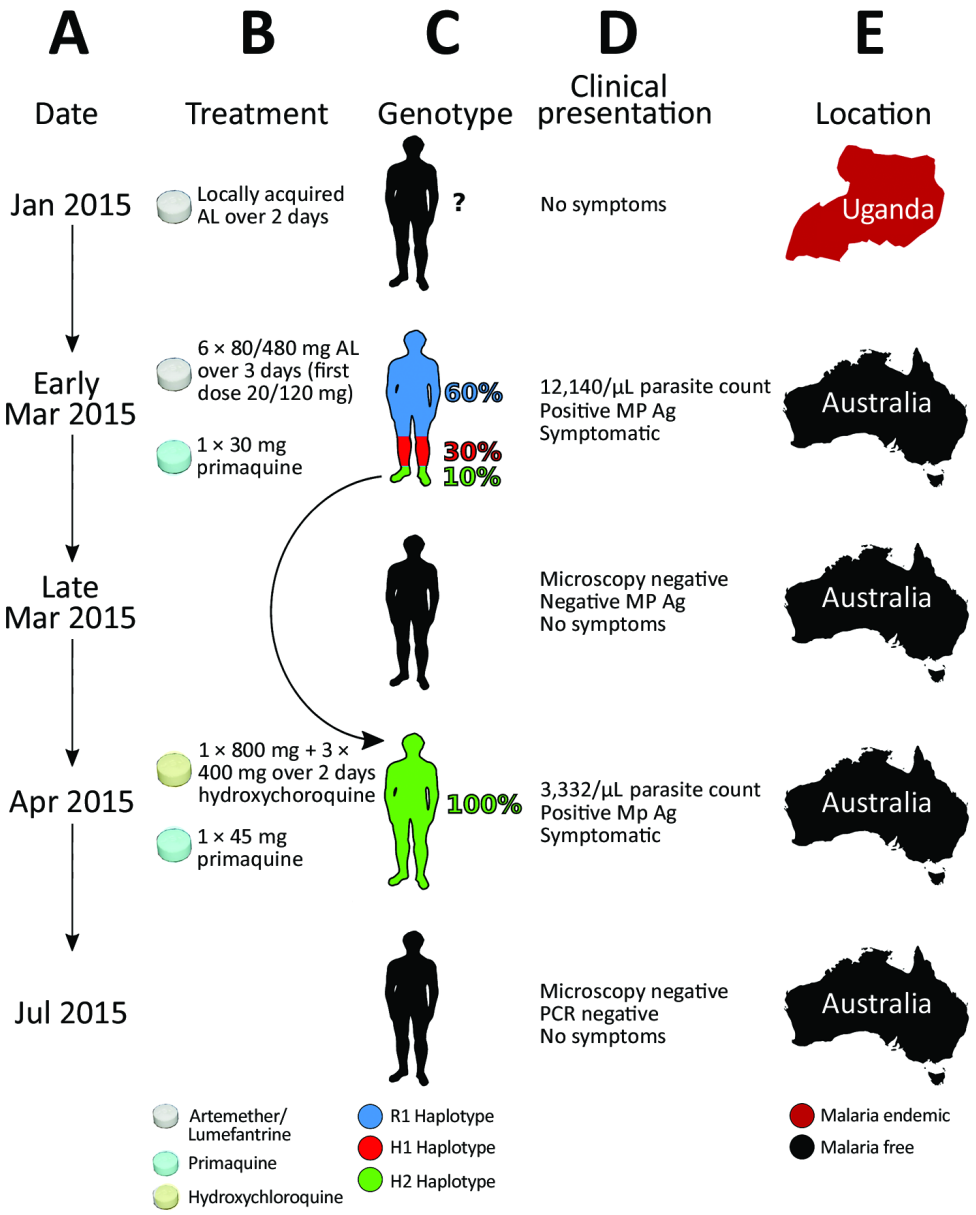


Figure 2.1: Timeline of the clinical case of a patient with Plasmodium malariae infection diagnosed and treated at Royal Darwin Hospital, Darwin, Northern Territory, Australia, March–April 2015, showing the timing (A), treatment (B), parasite's genotype as inferred from whole-genome sequencing (C), clinical presentation (D), and location (E). The rounded arrow indicates the recrudescence of the minor haplotype 2 in the initial infection to dominate monoclonally in the second infection. AL, artemether/lumefantrine; H1, haplotype 1; H2; haplotype 2; MP Ag, pan-malarial antigen; R1, reference haplotype.

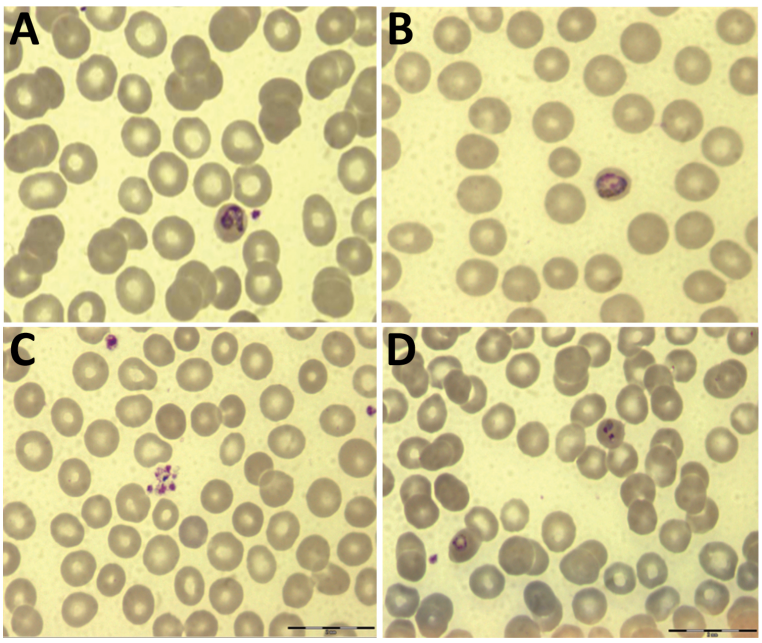


Figure 2.2: Positive Plasmodium malariae thin blood films. Thin smear scans of the initial infection (A, B) and of the re-crudescence (C, D), both indicating a P. malariae infection.

test results were negative.

The patient remained in urban Darwin but returned to the hospital 52 days later, on April 22, 2015, with a 2-week history of fevers, fatigue, and headache. Microscopy again identified *P. malariae* with a parasite count of 3,332/μL (figure 2.2 c & d). Chloroquine was unavailable, so the patient was retreated with oral hydroxychloroquine with an 800-mg loading dose, followed by 400 mg at 6 hours, 400 mg at 24 hours, 400 mg at 48 hours, and a single 45-mg dose of oral primaquine. The parasite count declined rapidly to 37/μL at 28 hours, 191/μL at 49 hours, and 76/μL at 88 hours of treatment. His symptoms resolved rapidly. Thick and thin blood films were negative on day 4 and remained negative on retesting at days 8, 35, 41, and 84, and the patient remained free of symptoms throughout. A PCR on blood collected at 12 weeks was also negative.

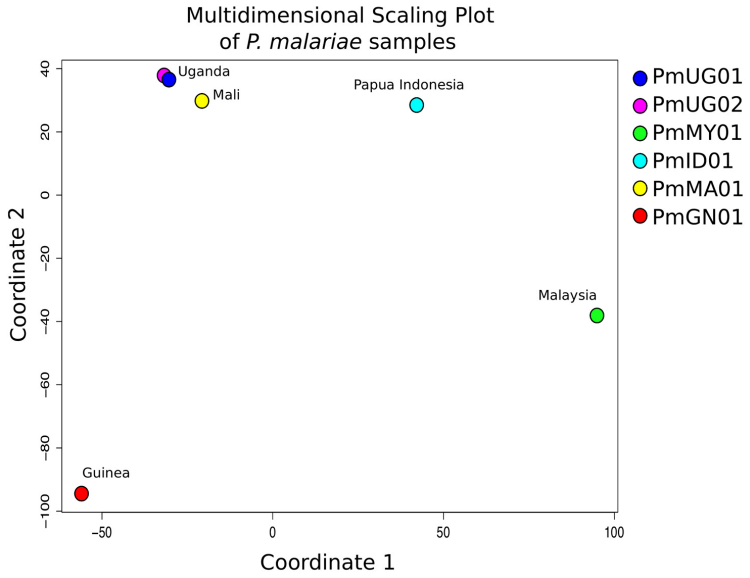


Figure 2.3: Similar SNP spectra for PmUG01 and PmUG02. Multidimensional scaling plot based on differences in SNP spectra between the different P. malariae samples [301], showing that the initial (PmUG01) and the recrudescent (PmUG02) infections are significantly more similar to each other than to the other P. malariae samples. This suggests that the two infections have a similar origin.

2.4 Whole-Genome Sequencing

Extensive sequencing was performed from blood samples obtained from the initial (PmUG01) and recrudescent (PmUG02) infection (appendix table C.1), covering >99% of the genome at >20x for both infections. By using additional *P. malariae* samples published previously [301], I identified single-nucleotide polymorphisms (SNPs) using GATK's UnifiedGenotyper (Broad Institute, Cambridge, MA, USA) [215] and filtered them based on several parameters (appendix table C.2). A multidimensional scaling plot of the samples based on their SNP allele frequency-spectra revealed that PmUG01 and PmUG02 were more closely related to each other than to any of the other samples (figure 2.3), as expected if they were related recombinants derived from the same original infection.

Searching solely for SNPs that distinguish PmUG01 and PmUG02, I identified 2,631 variants af-

Table 2.1: Changes in genotype calls between the two infections

Initial Genotype	Recrudescence Genotype	Number of Sites
HR	HA	1
h	h	217
h	HR	1,178
h	HA	1,038
HA	h	1*
HR	h	196
HA	HR	0

HR = Homozygous reference, HA = Homozygous alternate, h = heterozygous
* in repetitive region of rhoptry-associated membrane antigen

ter filtering (appendix table C.3). PmUG01 was the sample from which the reference genome (R1) was constructed[301], and only one SNP in PmUG01 suggested a nucleotide base different from the reference strain, probably because it was in a repetitive region (table 2.1). PmUG01 appeared to be a polyclonal infection with a bimodal distribution of alternate (*ie.*, nonreference) alleles at frequencies of 0.15 and 0.35 (figure 2.4 a). Conversely, PmUG02 appeared to be a monoclonal infection with 85% of sites being either fixed for the reference allele or for an alternative allele (figure 2.4 b). Comparison of the initial and recrudescent infections revealed that heterozygous sites in the initial infection had become either homozygous alternate (40%) or homozygous reference (45%) (table 2.1). Analysis of the genotype calls across the genome (figure 2.5) revealed that, whereas the heterozygous sites from the initial infection were spread evenly across the 14 chromosomes, the homozygous alternate sites in the recrudescent infection were present in distinct clusters, implying that the initial infection was polyclonal and that the recrudescence was attributable to a single clone that was closely related to the reference clone.

Comparison of the distribution of the alternate allele frequencies throughout the genome of the initial and recrudescent strain (figure 2.6) revealed bands of alleles at frequencies of 0.15 and 0.35 in

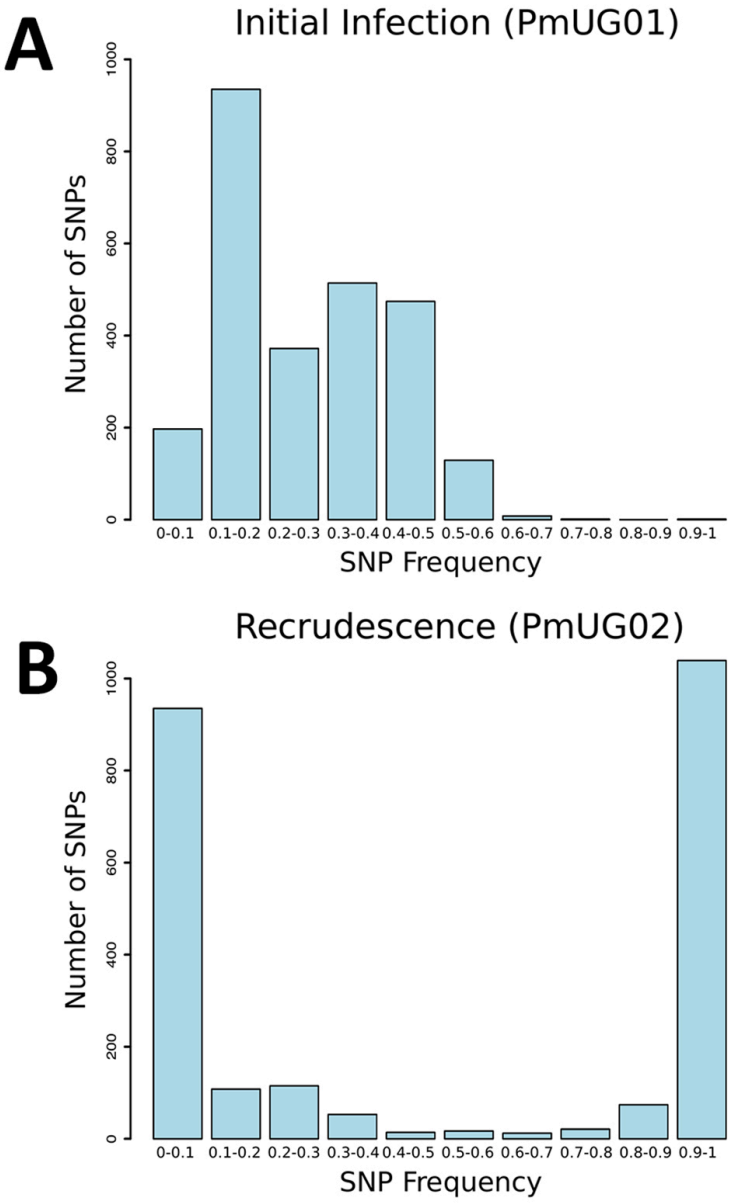


Figure 2.4: Differences in SNP frequencies between PmUG01 and PmUG02. SNP frequency bar plots for both the initial infection (A) and the recrudescence (B), showing that there was a significant shift in the SNP frequency spectra between the two infections, with the initial infection being a polyclonal infection, while the recrudescence seems to be monoclonal. Interestingly, the initial infection seems to have a bimodal distribution of heterozygous SNPs.

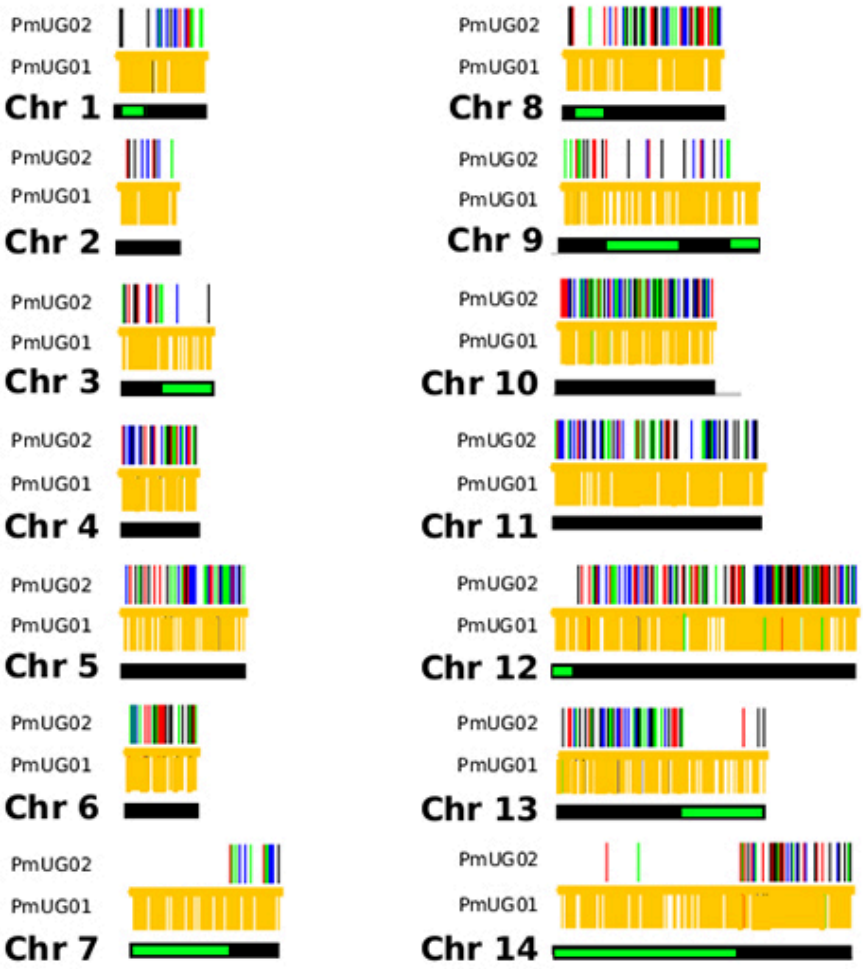


Figure 2.5: Clustered SNP distribution in PmUG02 across chromosomes. Distribution of heterozygous sites (yellow) in the initial infection (PmUG01) and homozygous alternate sites (other colors) in the recrudescence (PmUG02) across the 14 chromosomes of P. malariae. The different colors for the homozygous alternate SNPs are arbitrary. Chromosome regions with more heterozygous sites in the initial infection becoming homozygous reference than becoming homozygous alternate in the recrudescence are marked in green. Genotypes were plotted using Artemis [300].

the initial infection spatially clustered throughout the genome (figure 2.7). The alleles that increased in relative frequency in PmUG02 were mostly at frequencies of 0.15, whereas the alleles at frequencies of 0.35 decreased in frequency and the positions became homozygous reference in PmUG02 (figure 2.7 & 2.6). These data strongly suggested that, in addition to R1, two minor clones (minor haplotypes) were also present. Of these two, the clone with the haplotype comprising alternate alleles at frequencies of 0.35 (H1) appeared to have been eliminated during the drug treatment because no alleles specific to H1 were present in the recrudescent infection. The other minor clone comprised a haplotype with alternate alleles at frequencies of 0.15 (H2) in the initial infection; this clone appeared to have caused the recrudescence (figure 2.1 c). Based on the relative alternate allele frequencies of the 3 haplotypes in the initial infection, 60% of the parasites were of the R1 haplotype, 30% of H1, and 10% of H2. These estimates were broadly consistent with the ratio of alleles in tri-allelic sites (0.69:0.22:0.09) (appendix table C.4). The ratio of alleles in these tri-allelic sites changes markedly in PmUG02 (0.13:0.06:0.81), with over half of sites becoming homozygous for H2 but with some heterogeneity in the other sites (appendix table C.5), probably because of the low coverage depth and because they were in repetitive regions.

Unexpectedly, several SNPs at high allele frequencies (>0.4) also increased in frequency in the recrudescent strain. Testing by using additional *P. malariae* samples[301] showed that 80% of these SNPs were homozygous for the alternate allele in >1 other *P. malariae* samples, whereas 30% were homozygous in all other *P. malariae* samples (figure 2.8). This observation indicated that of these unusual SNPs, 50% were highly polymorphic, whereas 30% were probably low frequency SNPs with rare variants present in the reference strain. This would explain the observation of SNPs with high reference allele frequency in the initial infection that became homozygous alternate in the recrudescence, given

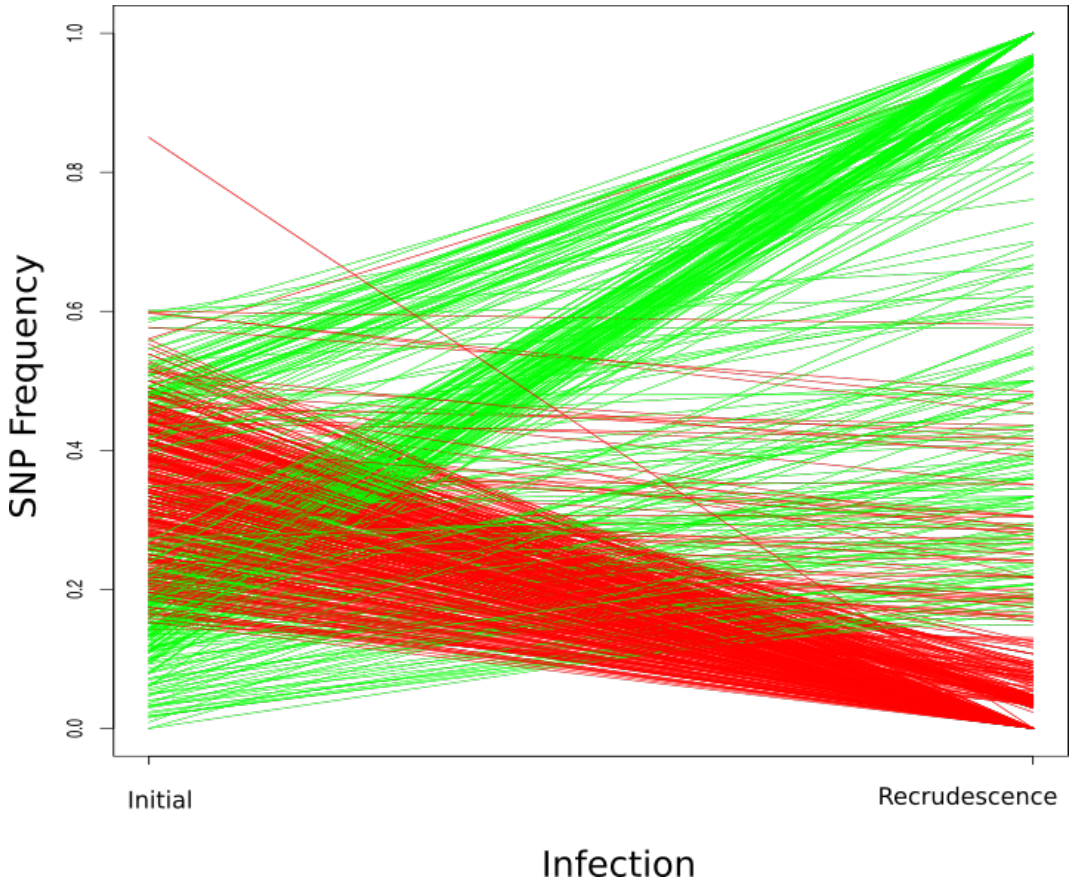


Figure 2.6: Changes in the SNP frequencies from the initial infection (left) to the recrudescence (right), coloured by whether the SNP increases in frequency (green) or decreases (red).

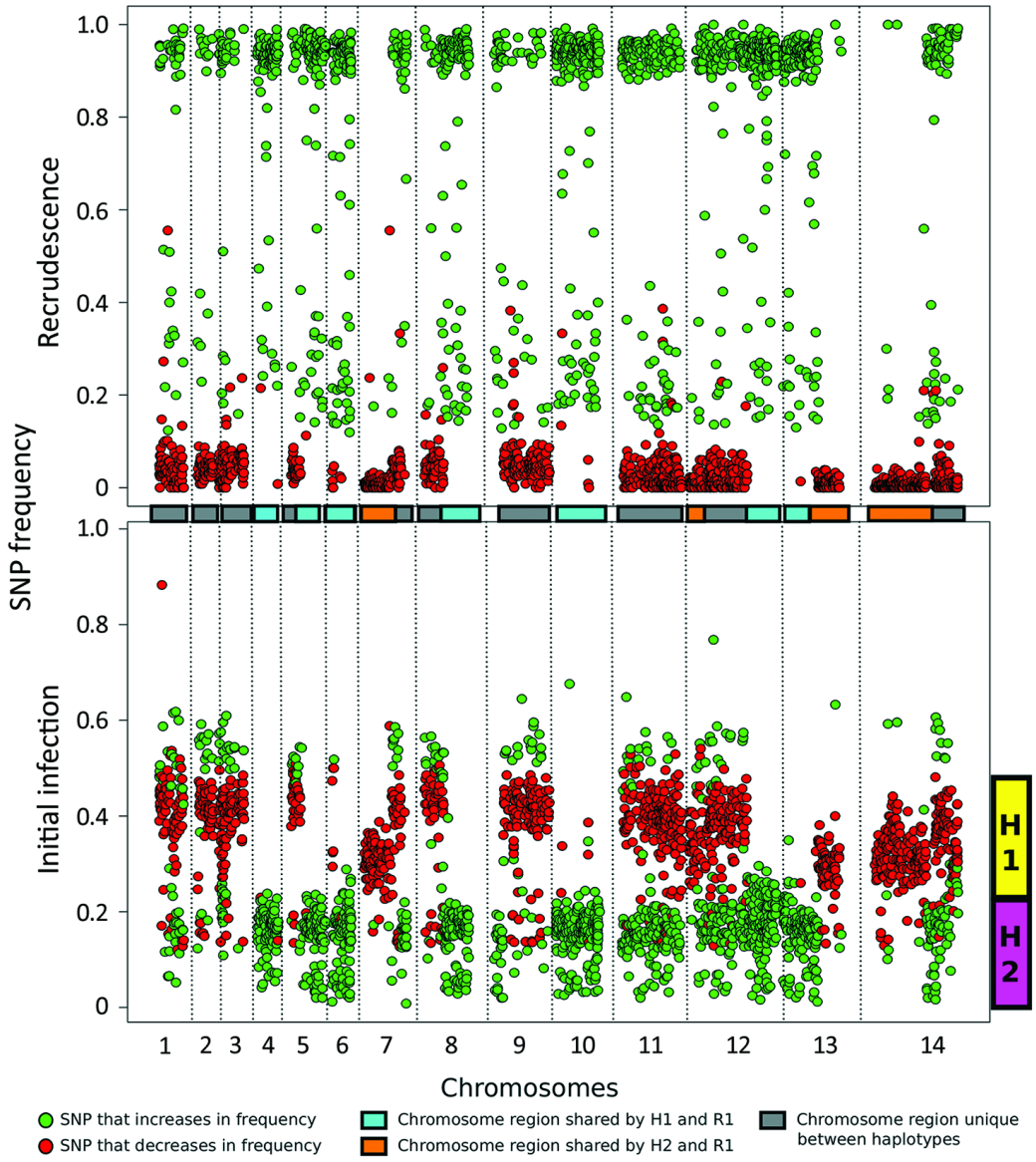


Figure 2.7: Analysis of the minor haplotype (H2) that caused recrudescence of Plasmodium malariae infection in a patient at Royal Darwin Hospital, Darwin, Northern Territory, Australia, March–April 2015, showing distribution of SNP alternative (nonreference) allele frequencies across the 14 chromosomes (boxes in the middle and dotted vertical lines) in the initial infection (bottom plot) and the recrudescence (top plot). The SNP colors (green, increasing in frequency; red, decreasing in frequency) form two clear bands, corresponding to H1 (yellow box) and H2 (pink box). H2 probably caused the recrudescence given that all of its alleles increase considerably in frequency. Colored boxes in center of chart indicate chromosome sharing. H1, haplotype 1; H2, haplotype 2; R1, reference genome; SNP, single nucleotide polymorphism.

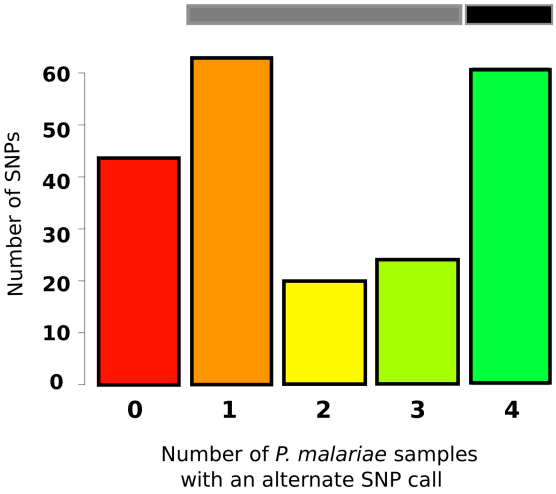


Figure 2.8: Unusual SNPs in other P. malariae samples. Presence/absence in other P. malariae samples of SNPs with SNP frequencies of over 0.4 in the initial infection that increased in frequency in the recrudescence. The black bar indicates SNPs that are present in all other P. malariae, suggesting that the reference strain is rather the variant compared to the general population. The gray bar indicates SNPs that are present in multiple other P. malariae samples, suggesting that they are highly polymorphic sites. In sum, this suggests that most of these SNPs are likely SNPs shared by both H1 and H2, explaining why they have high frequencies in the initial infection and in the recrudescence.

that they were probably SNPs with alternate alleles shared by H1 and H2.

To clarify the relationships of the different haplotypes with each other, I classified every genome region by whether any of the 3 haplotypes were identical to each other (figure 2.7). Approximately 25% of the genome is shared between H1 and R1 and between H2 and R1. No regions were shared between H1 and H2, which suggested that both H1 and H2 were half-siblings of R1, although they did not share any parent between themselves (figure 2.9). The finding that all haplotypes were related to each other through R1 further suggested that all strains were transmitted from the same mosquito bite and that the mosquito ingested at least 4 different parental haplotypes (figure 2.9).

Analysis of SNPs in orthologs of known drug-resistance genes identified three nonsynonymous SNPs in the multidrug resistance protein 2 (*pfmdr2*) gene, one of which was in the ABC transporter domain, and two in the ABC transporter domain of ABC transporter C family member 2, present

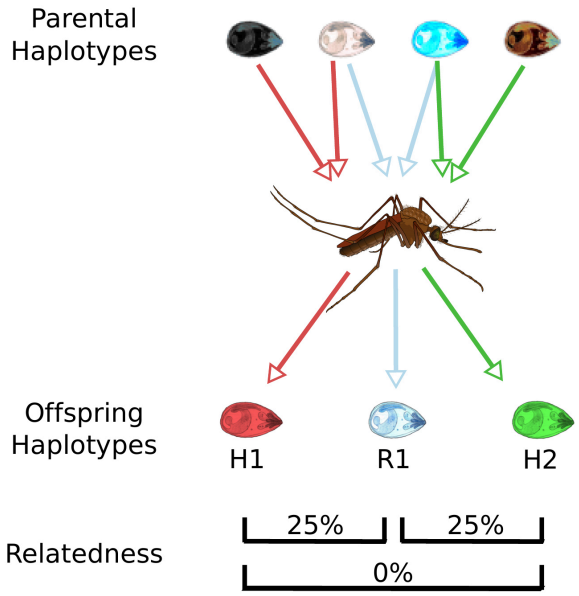


Figure 2.9: Inferred number of parental haplotypes. The relatedness of the three haplotypes in the initial infection (offspring haplotypes) as inferred by the sharing of genomic regions. This sharing suggests that there were four parental haplotypes present in the mosquito that interbred to form the three haplotypes we see in the initial infection. Of these, it seems that R1 is a half-sibling with both H1 and H2, but via a different parental haplotype.

in the recrudescent strain (H2) but not the other strains (table 2.2). No evidence was found for copy number variation in any gene compared with the reference strain, and the reference strain did not appear to have an amplification of the multidrug resistance protein 1 gene compared with any of the other *P. malariae* samples.

2.5 Discussion

This report of a case of recurrent *P. malariae* malaria is unusual in that it describes the molecular characterization and confirmation of a treatment failure after directly observed, appropriately administered, quality-assured AL dosing in a nonendemic environment where reinfection was not possible. Whole genome sequencing demonstrated that the recrudescence was attributable to a minor clone present in the initial polyclonal infection. The case raises two important questions: first, what was the cause of treatment failure; and second, why did recrudescence arise from the minor clone rather than a dominant reference clone?

Although the efficacy of AL for *P. malariae* infection is assumed in many national guidelines[176], *P. malariae* monoinfections are relatively unusual and often of low density. To my knowledge, there have been no published efficacy series of AL with the long follow-up necessary to assess efficacy against a parasite with a 72-hour life cycle. In a non-randomized efficacy study of 4 PCR-confirmed *P. malariae* infections treated with AL in Gabon (one *P. malariae* monoinfection and three mixed *P. malariae*/*P. falciparum* infections), all 4 were microscopy negative at day 28, with no follow-up beyond this time[223]. Among 80 PCR-confirmed *P. malariae*/*P. falciparum* mixed species infections in Uganda, 12% were still PCR-positive for *P. malariae* at day 7 and 6% were still PCR-positive on day 17[31]. An additional three reports have documented *P. malariae* infections occurring at 38 days, 47 days,

Table 2.2: H2 specific nonsynonymous and synonymous mutations in putative drug resistance genes

Gene ID	Gene Description	*P. falciparum* ortholog	nsSNPs	sSNPs	Resistance
01020700	Chloroquine resistance transporter	0709000	0	0	Chloroquine, Quinine
10021600	Multidrug resistance protein 1	0523000	0	2	Multiple
12069100	Multidrug resistance protein 2	1447800	3	0	Multiple, Artemisinin
14053100	AP-2 complex subunit mu	1218300	0	0	Artemisinin
02011900	ABC transporter C family member 1	0112200	0	0	Artemisinin, Sulfadoxine/Pyrimethamine
14063400	ABC transporter C family member 2	1229100	2	0	Multiple
02017400	Calcium-transporting ATPase	0106300	0	0	Artemisinin
13021900	P-type ATPase 4	1211900	0	1	Multiple
12021200	Kelch protein k13	1343700	0	0	Artemisinin
14020100	Sodium/hydrogen exchanger 1	1343700	0	0	Quinine
05034700	Bifunctional dihydrofolate reductase-thymidylate	0417200	2	1	Pyrimethamine
14036800	ATP-dependent Clp protease adaptor protein	0810800	0	0	Sulfadoxine
MIT001100	Cytochrome b	mal_mito_3	0	0	Atovaquone

and four months after AL treatment of an initial microscopy-diagnosed *P. falciparum* infection in returned travelers with no further possible malaria exposure [327,107,46].

Several plausible explanations might account for a recurrence of *P. malariae* parasitemia after treatment with AL (figure 2.10). The last indigenous case of malaria in the Northern Territory was in 1962, with no subsequent cases of introduced malaria or autochthonous transmission [170]; hence, the possibility of reinfection can be excluded (figure 2.10 a). Additionally, the presence of the H2 haplotype in the initial infection and recrudescent infection confirms treatment failure.

Inadequate drug absorption resulting in suboptimal serum drug concentrations can cause treatment failure (figure 2.10 b). Artemether is rapidly absorbed and eliminated (half-life of a few hours), whereas lumefantrine is variably absorbed and more slowly eliminated (half-life approximately 3.2 days) [99]. Lumefantrine is a lipophilic compound with erratic bioavailability unless administered with a small fatty meal [17], and for this reason, guidelines recommend administration of AL with a fatty meal such as milk or a small biscuit. In the case of this patient, adequate serum concentrations of lumefantrine could not be confirmed; however, the patient took a complete course of treatment, and all doses were supervised in the hospital and administered with a milk biscuit to ensure good absorption. None of the treatment doses were vomited. In this scenario, one would expect >98% efficacy against *P. falciparum* [376]. In addition, the clones associated with the R1 and H1 haplotypes, accounting for 90% of the parasite load, were cleared, suggesting that the plasma drug concentrations were sufficient to eliminate both infections. Nevertheless, considerable inter-individual variation exists in lumefantrine exposure, and this patient may have had relatively low concentrations.

Cure of malaria in a nonimmune patient requires that antimalarial blood concentrations are sustained above the parasites' miniumum inhibitory concentration (MIC) until the entire parasite biomass

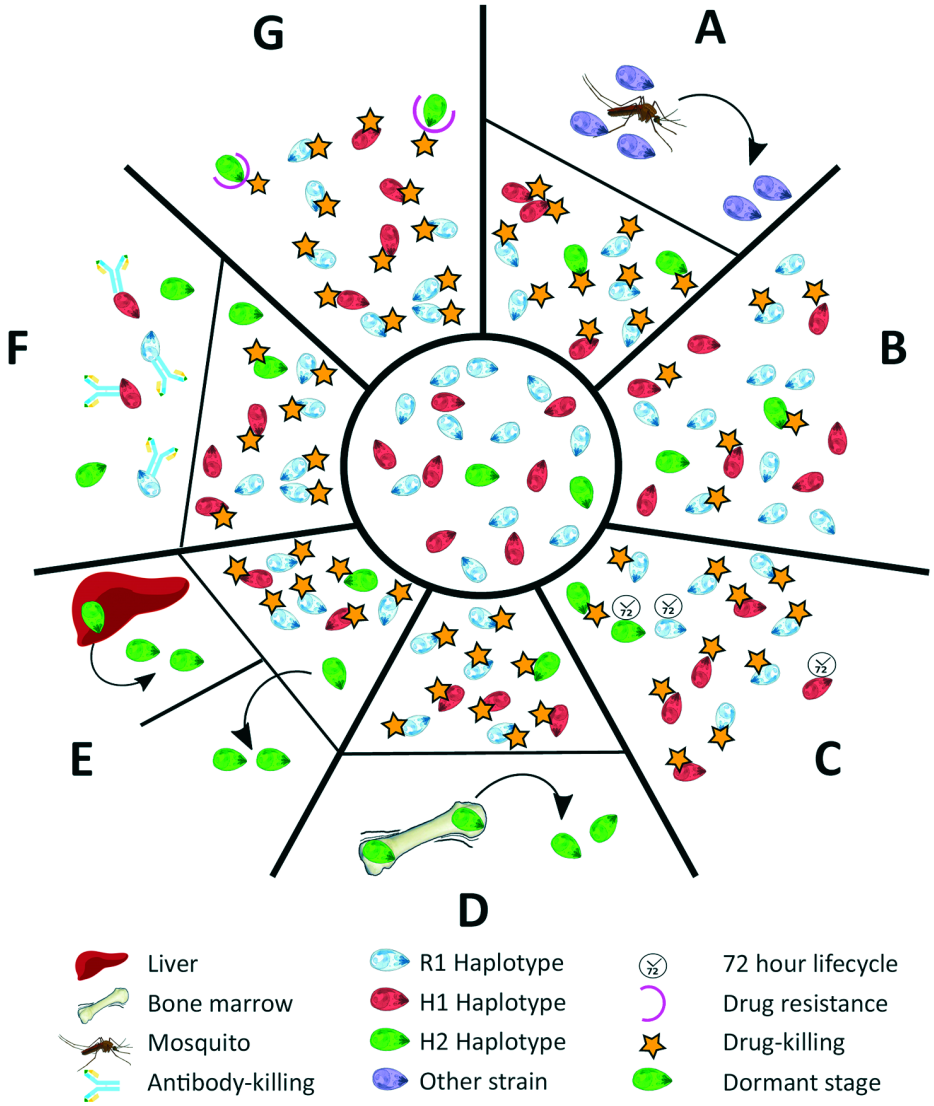


Figure 2.10: The different scenarios under which a second Plasmodium malariae infection could have occurred from the initial infection diagnosed in a patient at Royal Darwin Hospital, Darwin, Northern Territory, Australia, March–April 2015. Initial infection is shown in the inner circle. A) A completely new infection might have caused the second malaria onset. B) The drug might not have been absorbed at sufficient levels to kill all the parasites in the blood (pharmacokinetic cause). C) The longer intraerythrocytic lifecycle of P. malariae (72 hours) might have enabled some parasites to survive the drug action until lumefantrine concentrations became subtherapeutic (pharmacokinetic cause). D) H2 parasites might have differentially sequestered with a biomass out of proportion with the peripheral parasitemia. E) Some parasites might have formed dormant stages in the liver, blood, or elsewhere (pharmacodynamic cause). F) An immune response might have been differentially primed against haplotypes at higher biomass. G) A haplotype within the initial infection might have been relatively drug resistant (fitness advantage). H1, haplotype 1; H2; haplotype 2; R1, reference genome.

has been eliminated. In the presence of antimalarial drugs, the parasite biomass generally decreases over time in an exponential manner, with drug concentrations needing to be sustained above the MIC for >4 lifecycles[367]. In the case of this patient, the baseline parasitemia at initial presentation was 12,000/μL, which is relatively high compared with most *P. malariae* clinical infections[68]. Thus, the combination of the long parasite lifecycle resulting in one rather than two asexual cycles being exposed to artemether, and the short period (16 days) for which lumefantrine was at concentrations sufficient to kill the parasite may have resulted in parasites surviving the initial treatment and reestablishing a chronic parasitemia that was then sustained for 50 days before recrudescing (figure 2.10 c).

Another possibility is that some parasites could have sequestered (figure 2.10 d) or become dormant (figure 2.10 e). Whereas dormancy would allow a proportion of the parasites to evade blood stage schizontocidal activity, plausible sites for sequestration of *P. malariae*–infected erythrocytes would still be exposed to therapeutic concentrations of blood stage antimalarials, making this possibility an unlikely explanation for this patient's recrudescent infection. *P. malariae* is well-recognized as having a prolonged preerythrocytic phase and a prepatent period of 16–59 days[68]. The initial treatment course of AL was administered 56 days after the patient left Uganda, so any preerythrocytic stages were probably not present at the time of initial AL treatment.

Although the ability to form hypnozoites (dormant exoerythrocytic stages) occurs in three human malaria parasite species (*P. vivax*, *P. ovale curtisi*, and *P. o. wallikeri*), the evidence that latent exoerythrocytic stages do not occur in *P. malariae* is limited[68]. Case reports have documented *P. malariae* producing symptomatic disease many years after exposure to infection, as noted in the case of a 74-year-old woman in Greece with *P. malariae* reactivation after >40 years[357]. Such latency suggests that low-level parasitemia could persist for many years after infection, and indeed may be lifelong. In the

case of this patient, parasite recrudescence occurred >100 days after he had left a malaria-endemic area.

Although inadequate drug absorption (figure 2.10 b), duration of treatment (figure 2.10 c), or dormancy (figure 2.10 e) all may have contributed to parasite recrudescence, these indiscriminate explanations would be expected to occur primarily in the dominant strain during the initial infection [367] (figure 2.11 a). One could speculate that the H2 minor parasite population might have emerged from a hepatic schizont that ruptured days after those giving rise to the majority haplotypes and, despite genetic similarity, had substantial differences in surface antigenicity. The antibody response to the primary infection, which would have reached a maximum 3 weeks after the illness began, would have been directed against the majority haplotypes and might not have recognized the minor population (figure 2.10 f). Alternatively, more of the minor population might have been in the dormant state compared with the dominant circulating clones with R1 or H1 haplotypes, or more might have been at a higher biomass in erythrocytes sequestered elsewhere, enabling a proportion to evade antimalarial drug action and recrudesce.

Finally, the minority clone with the H2 haplotype might have recrudesced because of a fitness advantage over the other clones/haplotypes (figure 2.11 b & c), possibly including relative resistance to either artemether or lumefantrine (figure 2.10 g). In *P. falciparum*, resistance to artemether is acquired through mutations in the propeller domain of K13 [334], whereas *P. falciparum* resistance to lumefantrine is associated with mutations and copy number variation in the *pfmdr1* gene [283,356]. Although neither of these genes had nonsynonymous mutations in H2, one nonsynonymous mutation was noted in the ABC transporter domain of *pfmdr2*, potentially involved in artemisinin resistance [221,208], and two nonsynonymous mutations were noted in the *pfmrp2* gene, which has been implicated in reduced *ex vivo* susceptibility to lumefantrine in *P. falciparum* [251]. I also identified two

nonsynonymous SNPs in the dihydrofolate reductase homologue. Low serum concentrations, a modest reduction in lumefantrine efficacy, the prolonged life cycle of *P. malariae*, and rapid elimination of lumefantrine all might have contributed to the observed treatment failure in this patient.

2.6 Conclusion

I have described a case of *P. malariae* recrudescence occurring in a non–malaria-endemic country after adequately administered AL. Whole-genome sequencing data revealed that the monoclonal recrudescence consisted of a minor haplotype that accounted for ≈10% of the initial infection and that all the haplotypes in the initial infection were related to each other and therefore probably originated from the same infective mosquito bite. Although the haplotypes were closely related, the genomic data suggests that 4 parental haplotypes were ingested by the mosquito, indicating considerable diversity and transmission of *P. malariae*. This case raises concerns about the adequacy of ACTs with a short half-life partner drug, such as AL, in treating *P. malariae* infections and suggests that optimal ACTs to treat *P. malariae* should include a slowly eliminated partner drug. This reinforces the importance of a longer duration of follow-up monitoring of patients infected with *P. malariae* for late recrudescence.

The newly assembled *P. malariae* reference genome sequence opened up the possibility of analysing and better characterising this particular case of recrudescent malaria. While it was possible to gain new insight into the potential sequence of events that led to the present case of *P. malariae* recrudescence, little is known about drug resistance mechanisms in *P. malariae* and consequently only inferences based on sequence homology could be made. Even for *P. falciparum*, understanding the genetic basis of drug resistance is not straightforward. In Chapter 3, I will present an analysis I performed to better

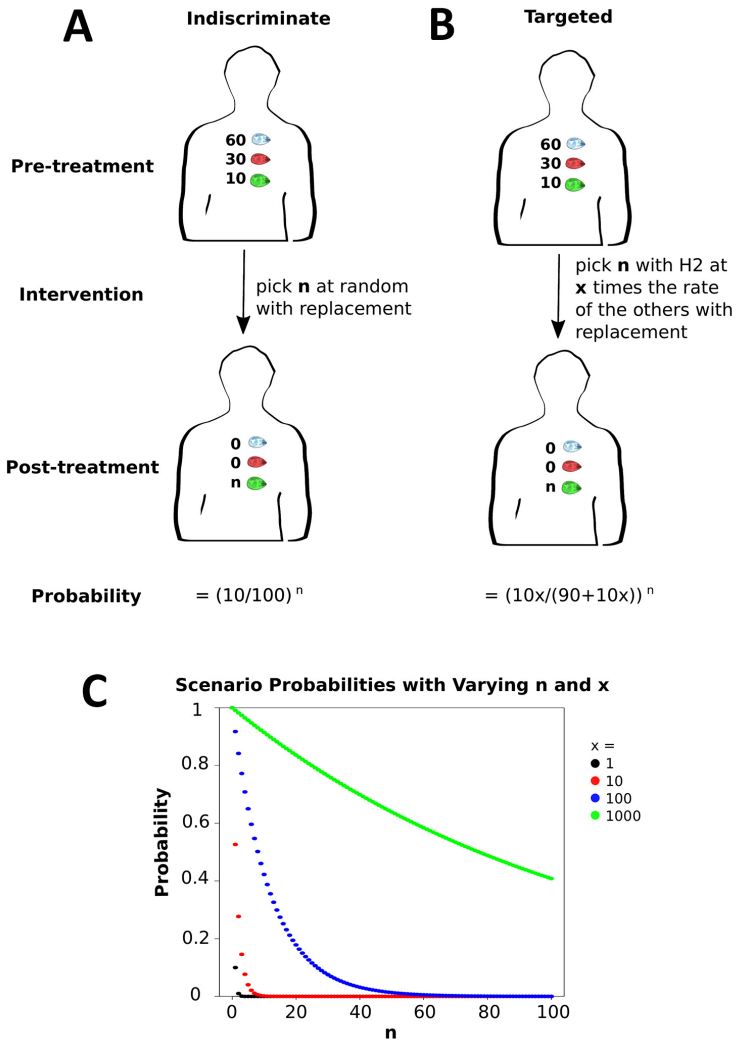


Figure 2.11: Indiscriminate versus targeted interventions. Differences in probabilities for (A) indiscriminate interventions (i.e., explanations for a recrudescence) that affect all haplotypes in the infection equally, such as insufficient drug dosage or drug avoidance through a longer lifecycle, versus (B) targeted interventions that potentially affect one haplotype different to another, for example haplotype-specific drug resistance or an increased propensity for greater sequestered biomass in one haplotype. The two scenarios show that the difference is in how an intervention 'selects' for n number of parasites in the recrudescence. In an indiscriminate intervention, all haplotypes have the same probability of being selected, while H2 is x times more likely to be selected in a targeted intervention. An indiscriminate intervention has a low probability (<0.05) at all values of n > 1 (C), while the probability of a targeted intervention increases across all values of n the higher x. If an indiscriminate intervention were to be the sole explanation for a recrudescence, then it would suggest that only a singly parasite survived from the entire initial infection, an unlikely scenario. On the other hand, a targeted intervention presents a more parsimonious explanation for the lower level haplotype recrudescing.

understand the genetic basis of mefloquine response in *P. falciparum* malaria parasites, specifically those that have acquired multidrug resistance in the Southeast Asian setting.

Science, especially natural and medical science, is always undergoing evolution, and one can never hope to have said the last word upon any branch of it.

Alphonse Laveran, Arch. Méd. Expérim., 1892 CE

3

Genetic architecture of mefloquine sensitivity in KEL1/PLA1 *P. falciparum*

3.1 Abstract

The current frontline antimalarial combination therapy, dihydroartemisinin-piperaquine, is rapidly losing efficacy in Southeast Asia as the multidrug resistant KEL1/PLA1 *P. falciparum* co-lineage spreads through the region. In the field, the spread of KEL1/PLA1 parasites has been accompanied

by the disappearance of the *mdr1* copy number amplification, associated with mefloquine resistance. This suggests that artesunate-mefloquine is an effective replacement first-line treatment. However, the speed at which *mdr1* copy number has changed raises questions about the sustainability of a switch to a mefloquine-based treatment. In this chapter I present an analysis of 430 Cambodian *P. falciparum* field isolates that were whole-genome sequenced and assayed *in vitro* for mefloquine response (IC_{50}). I show that KEL1/PLA1 parasites are hypersensitive to mefloquine as a result of having merged with the PLA1 lineage. Furthermore, a genome-wide association analysis identified a novel SNP associated with mefloquine sensitivity, which has rapidly increased in frequency from 2010 to 2013. The mutant allele at this SNP is only observed to occur on the hypersensitive KEL1/PLA1 background, further increasing the sensitivity to mefloquine of these parasites. The combined effect of the deamplification of *mdr1*, the acquisition of the PLA1 lineage, and the emergence of this novel SNP have resulted in a complex genetic architecture of mefloquine sensitivity, which suggests that artesunate-mefloquine could be effective and sustainable in the field.

3.2 Introduction

Artemisinin combination therapies, consisting of a potent but short-acting artemisinin derivative and a longer-acting partner drug, are the frontline treatment for severe and uncomplicated malaria [368]. Dihydroartemisinin-piperaquine treatment, used in Cambodia, Vietnam, Thailand, Myanmar, China, and Indonesia, is now increasingly becoming ineffective as resistance to treatment has emerged [187,185]. Resistance to artemisinin, characterized by a reduction in the speed of parasite clearance, was first reported in 2008 in Western Cambodia [243], from which it quickly spread throughout the region [84,15,351]. While these slow-clearing artemisinin-resistant parasites were initially still cleared by the partner drug,

piperaquine, they eventually acquired resistance to it[9], leading to cases of complete treatment failure[185,187,6].

Both artemisinin and piperaquine resistance have a genetic basis, with artemisinin resistance having been linked to mutations in the propeller domain of the *kelch 13* gene[219,221,13], and piperaquine resistance being associated with a copy number amplification of the *plasmepsin 2* and *plasmepsin 3* genes[8,372]. While artemisinin resistance has been shown to have emerged multiple times in Southeast Asia[342], one particular lineage, KEL1, outperformed others and eventually combined with the PLA1 lineage, associated with piperaquine resistance[9]. This multidrug resistant KEL1/PLA1 lineage is the main agent of the current outbreak of multidrug resistance in Southeast Asia[9,150], having spread from Western Cambodia to other countries[149], including Thailand, Laos, and Vietnam[6,149,307,346].

In response to the loss of efficacy of dihydroartemisinin-piperaquine, alternative antimalarial drugs are increasingly being looked at, including artesunate-mefloquine[186]. Anecdotal reports from the field indicate that artesunate-mefloquine remains an effective treatment in regions where KEL1/PLA1 is widespread[186]. Mefloquine had been used as a monotherapy in the region during the late 1980s and early 1990s, which eventually led to the rise of mefloquine resistance[245]. The genetic basis of mefloquine resistance was mapped to a copy number amplification of the *mdr1* gene[370,70,282], which was widespread in Southeast Asia by the turn of the millennium[282,246]. As KEL1, and eventually KEL1/PLA1, increased in frequency, the frequency of the *mdr1* copy number amplification decreased in the region[9,149], potentially leading to the current situation of mefloquine being effective in the field.

As a result, deploying artesunate-mefloquine throughout the region is now an attractive option, however the rapid spread of KEL1/PLA1 and the corresponding deamplification observed for *mdr1*[9,149] put into question the sustainability of deploying mefloquine in the long term. This concern is further

exemplified by the recent observation of a rise in frequency of triple mutants in the region[297]. As a consequence, a number of questions remain before this course of action can feasibly be considered. While artesunate-mefloquine appears to be successful in regions where KEL1/PLA1 is widespread, the individual response KEL1/PLA1 to mefloquine is not fully understood. If KEL1/PLA1 is mefloquine sensitive, then is this due to the loss of the *mdr1* amplification or does KEL1/PLA1 exhibit genetic changes that actively contribute to the sensitivity? Are there any other genomic changes in the parasite population that associate with this observed mefloquine sensitization? And finally, is there any evidence to believe that this sensitization is sustainable and that triple resistance won't arise?

In this chapter, I present an analysis of 430 Cambodian *P. falciparum* field isolates that were whole genome sequenced and assayed *in-vitro* for mefloquine (MQ) 50% inhibitory concentration (IC_{50}). I show that KEL1/PLA1 parasites appeared to be hypersensitive to mefloquine as a result of having merged with the PLA1 lineage, specifically through the acquisition of the *plasmepsin 2/3* copy number amplification. Furthermore, a genome-wide association study (GWAS) of all the samples identified a novel SNP (F1068L) in the *mdr1* gene that appears to associate with mefloquine sensitivity and has rapidly increased in frequency from 2010 to 2013. This SNP is only observed on the hypersensitive KEL1/PLA1 background, leading these parasites to become ultrasensitive to mefloquine. The deamplification of *mdr1*, amplification of *plasmepsin 2/3*, and the acquisition of this novel *mdr1* SNP result in a complex genetic architecture of mefloquine sensitivity that suggests that artesunate-mefloquine may be more effective and sustainable in the field than initially thought.

All methods used for this chapter are detailed in Appendix D. All sample collection, sequencing, and *in vitro* drug resistance assays were performed by collaborators. All data analysis presented in this chapter was performed by me. A similar analysis that I performed using chloroquine IC_{50} values is

described in Appendix E.

3.3 KEL1/PLA1 is hypersensitive to mefloquine

Originating from three provinces in Cambodia (Pursat, Preah Vihear, and Ratanakiri), 430 clinical *P. falciparum* isolates were collected between 2010 and 2013 in clinical studies[15,6,8,7] (table 3.1). At the time of collection, artemisinin and piperaquine resistance were entrenched in Pursat, emerging in Preah Vihear and rare in Ratanakiri[6,8,7,15], as mediated by the outbreak of the KEL1/PLA1 *P. falciparum* co-lineage from Pursat to Preah Vihear[9] and beyond[150]. To better understand the individual response of KEL1/PLA1 co-lineage parasites to mefloquine treatment, the collected samples were subjected to whole genome sequencing (WGS) and phenotyped for mefloquine (MQ) 50% inhibitory concentration (IC_{50}).

Table 3.1: Sample origin and collection year information

Year	Pursat	Preah Vihear	Ratanakiri	Total
2010	53	0	42	95
2011	76	58	54	188
2012	26	31	22	79
2013	37	19	12	68
Total	192	108	130	430

Number of samples whole genome sequenced and phenotyped from each of the provinces each year.

Of the 430 *P. falciparum* samples, 77 (17.9%) were KEL1/PLA1 parasites, 48 (11.2%) were KEL1/non-PLA1, 27 (6.3%) were non-KEL1/PLA1, and 268 (62.3%) did not belong to either lineage, with 10 (2.3%) samples being unclassifiable due to missing data. Additionally, *mdr1* amplifications were present across these groups (figure 3.1). I investigated the average response to MQ treatment of these different parasite populations stratified by *mdr1* amplification, as samples with the amplification were

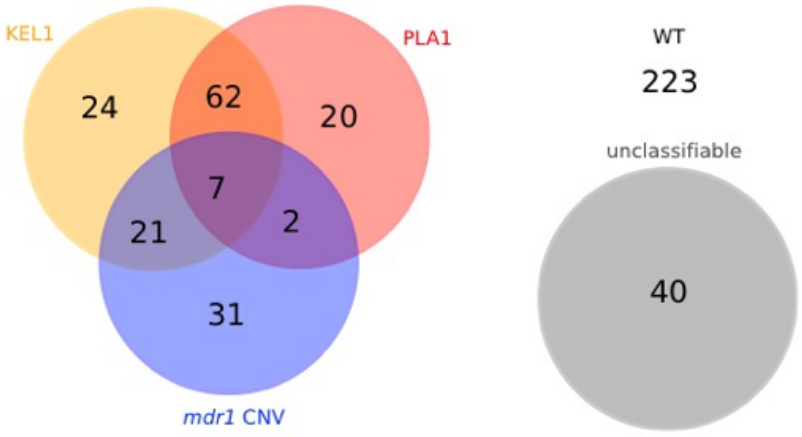


Figure 3.1: Showing the distribution of genotypes across the dataset, including KEL1, PLA1, and mdr1 amplification genotype. Samples for which any of these three genotypes cannot be ascertained are considered unclassifiable.

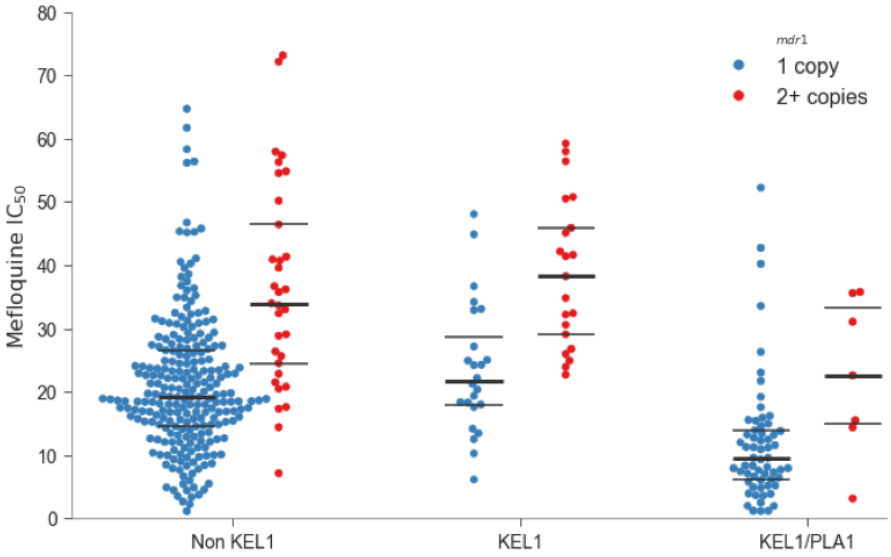


Figure 3.2: Mefloquine IC_{50} by KEL1/PLA1 lineage membership and stratified by mdr1 CNV. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

significantly more resistant to MQ than those without (mean IC_{50} ± SD: 35.4 nM ±15.0 vs. 19.5 nM ±11.1, Wilcoxon rank sum test with continuity correction: $p < 5 \times 10^{-15}$) (figure 3.2). I found that KEL1/PLA1 parasites appeared to be significantly more sensitive to mefloquine than wild type (WT) parasites (12.7 nM ±10.2 vs. 23.5 nM ±12.3, $p < 5 \times 10^{-15}$). This very significant difference remains true when I stratify both by samples without *mdr1* amplification (11.8 nM ±9.8 vs. 21.7 nM ±10.4, $p < 5 \times 10^{-14}$) and by those with the amplification (22.6 nM ±12.3 vs. 37.2 nM ±16.5, $p < 0.05$). While I find a significant difference between KEL1/non-PLA1 samples compared to WT parasites (30.6 nM ±13.1 vs. 23.5 nM ±12.3, $p < 5 \times 10^{-4}$), this difference is due to the overrepresentation of the *mdr1* amplification in the KEL1/non-PLA1 group compared to WT parasites (47% [21/45] vs. 12% [31/254]) and disappears when we control for *mdr1* amplification (*mdr1* amplification present: 38.7 nM ±11.7 vs. 37.2 nM ±16.5, $p = 0.52$; *mdr1* amplification absent: 23.6 nM ±10.4 vs. 21.7 nM ±10.4, $p = 0.30$). I do however find a significant difference between KEL1/non-PLA1 parasites and those of the KEL1/PLA1 co-lineage overall (30.6 nM ±13.1 vs. 12.7 nM ±10.2, $p < 5 \times 10^{-12}$), and when controlling for *mdr1* status (*mdr1* amplification present: 38.7 nM ±11.7 vs. 22.6 nM ±12.3, $p < 0.05$; *mdr1* amplification absent: 23.6 nM ±10.4 vs. 11.8 nM ±9.8, $p < 1 \times 10^{-6}$). These results strongly suggest that KEL1/PLA1 is hypersensitive to mefloquine, likely through the acquisition of the PLA1 component of the co-lineage.

3.4 The *MDR1* and *PLASMEPSIN 2/3* CNVs may be antagonistic

The PLA1 lineage is principally characterized by the presence of an amplification of the *plasmepsin 2* and *plasmepsin 3* genes[9], presence of which is linked to piperaquine resistance[8,372]. It has previously been shown that *mdr1* amplification has the opposite effect to the *plasmepsin 2/3* amplification,

namely increasing the piperaquine sensitivity of parasites with the *mdr1* amplification [8]. Additionally, using the Pf6 dataset (www.malariagen.net), a collection of over 5,000 *P. falciparum* WGS sampled globally, I observe that the two amplifications co-occurred less frequently in Cambodia than expected by chance (Fisher's exact test for count data: $p < 5 \times 10^{-6}$) (table 3.2). As a result of these apparent links between *mdr1* amplifications and *plasmepsin 2/3* amplifications, I investigated the interplay of the two types of amplification on MQ IC_{50} by looking at samples for which both amplification genotypes could be ascertained (figure 3.3). I found a very strong effect of the two copy number amplifications impacting the MQ IC_{50} in an opposing manner, with *mdr1* amplifications increasing MQ IC_{50} compared to WT parasites (37.8 nM $\pm$14.6 vs. 21.9 nM $\pm$10.4, $p < 5 \times 10^{-13}$) and *plasmepsin 2/3* amplifications decreasing MQ IC_{50} compared to WT (12.3 nM $\pm$10.1 vs. 21.9 nM $\pm$10.4, $p < 5 \times 10^{-16}$). There is consequently a very significant difference between parasites with an *mdr1* amplification and those with a *plasmepsin 2/3* amplification (37.8 nM $\pm$14.6 vs. 12.3 nM $\pm$10.1, $p < 1 \times 10^{-16}$). Interestingly, parasites containing both types of amplification appear to have an MQ IC_{50} phenotype comparable to that of WT parasites (23.0 nM $\pm$10.7 vs. 21.9 nM $\pm$10.4, $p = 0.52$), suggesting that the opposing effects of the two amplifications phenotypically cancel each other out. The observation that the two copy number amplifications appear to impact both mefloquine and piperaquine resistance in opposing manners appears to suggest that they may operate in an antagonistic manner. The fact that parasites with both amplifications are as susceptible to mefloquine as WT parasites, indicates that triple mutants [297] may not necessarily exhibit triple resistance.

Table 3.2: Copy number amplification co-occurrence

		plasmepsin 2/3 CNV		Total
		Absent	Present	
mdr1 CNV	Absent	438	198	636
	Present	112	14	126
Total		550	212	762

Pf6 samples from Cambodia with both genotypes ascertained and filtered for QC Pass.
Fisher's exact test for count data, $p < 5 \times 10^{-6}$

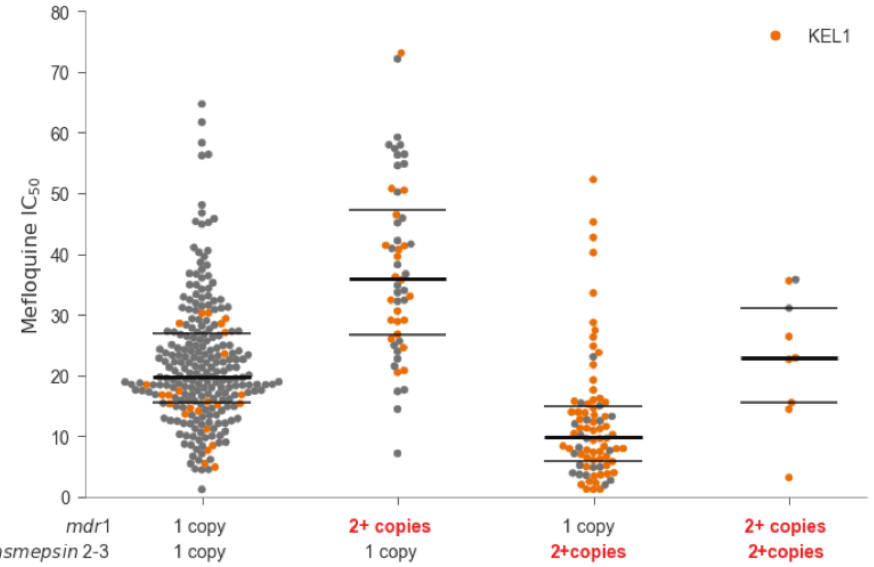


Figure 3.3: Mefloquine IC_{50} by presence and absence of the two copy number amplifications. Samples of the KEL1 lineage are highlighted in orange. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

3.5 A novel *MDR1* SNP associates with increased mefloquine susceptibility

In order to identify additional markers of mefloquine resistance other than the *mdr1* amplification, I conducted a genome-wide association study (GWAS) analysis using MQ $\log(IC_{50})$ as the dependent variable. I called SNPs across all 430 samples with MQ IC_{50} values and, following filtering (appendix D), retained 12,445 SNPs. Controlling for sample origin and for population structure (Appendix D), I identified 15 SNPs that passed the suggestive threshold ($p < 4 \times 10^{-4}$), one of which exceeded

the threshold for genome-wide significance ($p < 4 \times 10^{-6}$) (table 3.3 & figure 3.4). Unexpectedly, this particular SNP is located within the *mdr1* gene and codes for an F1068L amino acid alteration. Samples with the F1068L mutation appear to have a significantly lower MQ IC_{50} compared to wild type samples (6.5 nM ±5.2 vs. 22.4 nM ±12.9, $p < 5 \times 10^{-9}$), *ie.* samples with this mutation are more susceptible to mefloquine. Looking further into the 19 samples with this SNP, I noticed that all of them contain a *plasmepsin 2/3* amplification and no *mdr1* amplification (one sample had a missing genotype call for the *mdr1* CNV). Rerunning the GWAS analysis while controlling for both CNVs, yielded the same F1068L *mdr1* SNP as the most significant SNP genome-wide (figure 3.5 & table 3.5), suggesting that the SNP has an effect independent of the two copy number amplifications. Comparing the MQ IC_{50} of samples with the SNP with those without, while excluding samples with *mdr1* amplifications and stratifying by *plasmepsin 2/3* amplification, I found that the F1068L SNP associates with increased MQ sensitivity above and beyond the hypersensitivity induced by the *plasmepsin 2/3* amplification alone (6.5 nM ±5.2 vs. 13.9 nM ±10.6, $p < 5 \times 10^{-4}$) (figure 3.6), making these samples ultrasensitive to mefloquine.

3.6 The genetic architecture of mefloquine hypersensitivity

The F1068L SNP in *mdr1* is a novel SNP and has not previously been described in the literature. The SNP occurs in the 9th putative transmembrane domain of the MDR1 protein (www.uniprot.org), polymorphisms in which have previously been implicated with chloroquine resistance[254], suggesting that an amino acid alteration here may play a functional role. In the Pf6 dataset (www.malariagen.net), the SNP is limited to Cambodia and has not been previously observed in any other region. The SNP first appeared in 2010 with an allele frequency of 2% and then rapidly increased in frequency

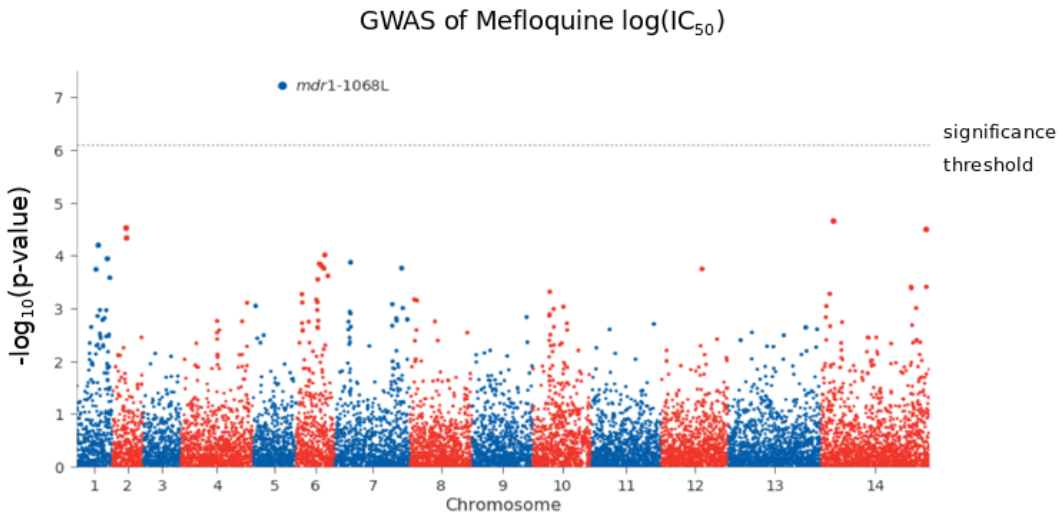


Figure 3.4: GWAS analysis using mefloquine log(IC_{50}) as the dependent variable. Each point corresponds to a SNP, coloured by the chromosome it is located on and they are ordered by their position on the chromosome. The line indicates the genome-wide threshold for significance ($p < 4 \times 10^{-6}$). The SNPs exceeding the significance threshold are labeled by the gene they are found in and the amino acid alteration they code for.

Table 3.3: SNPs most strongly associated with linear MQ IC_{50}

Chromosome	Position	Gene ID	Gene Description	N or S	Alteration	p-value
5	961,093	PF3D7_0523000	Multidrug resistance protein 1	N	F1068L	5.5×10^{-8}
6	909,326	PF3D7_0622300	Vacuolar transporter chaperone	N	S330F	8.4×10^{-6}
1	532,536	PF3D7_0113800	DBL containing protein	S		1.1×10^{-5}
6	831,420	PF3D7_0619800	Conserved Plasmodium protein	N	D2487N	1.2×10^{-5}
6	1,068,475	PF3D7_0626400	Sec14 domain containing protein	N	K1110R	1.2×10^{-5}
1	535,099	PF3D7_0113800	DBL containing protein	N	E2484G	1.3×10^{-5}
2	438,323	PF3D7_0210800	Conserved Plasmodium protein	N	K227R	1.3×10^{-5}
6	806,029	PF3D7_0619300	Conserved Plasmodium protein	N	S2842C	1.4×10^{-5}
6	703,002	PF3D7_0616900	Conserved Plasmodium protein	N	F1251I	1.9×10^{-5}
6	878,937	PF3D7_0621400	Pf77 protein	N	I50V	2.0×10^{-5}
2	463,586	PF3D7_0211500	GAF domain-related protein	N	S142G	2.2×10^{-5}
14	420,625	PF3D7_1410400	Rhoptry-associated protein 1	N	D62N	3.4×10^{-5}
7	304,718	PF3D7_0706200	Conserved Plasmodium protein	N	N66S	4.2×10^{-5}
6	664,920	PF3D7_0615900	Protein phosphatase	N	R979P	7.5×10^{-5}
1	531,949	PF3D7_0113800	DBL containing protein	N	E1526K	9.3×10^{-5}

Table 3.4: SNPs that pass either the Bonferroni-adjusted p-value threshold ($p < 4 \times 10^{-6}$) or the more lenient suggestive threshold ($p < 1 \times 10^{-4}$) are listed in order of increasing p-value. The table shows chromosome and nucleotide position of the SNP, the ID and description of the gene in which the SNP occurs, whether it is a synonymous (S) or nonsynonymous (N) mutation, what amino acid alteration it encodes if it is nonsynonymous, and the p-value associated with the SNP.

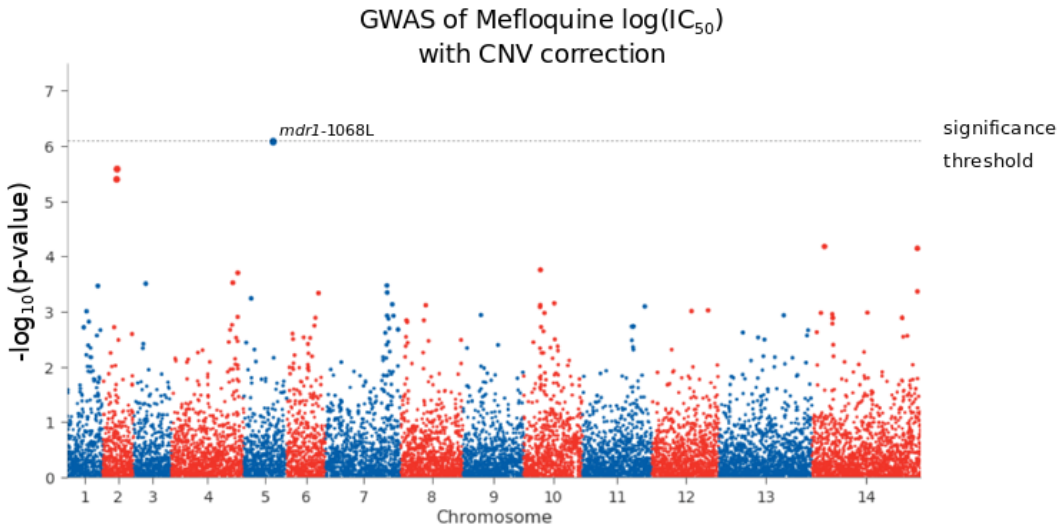


Figure 3.5: GWAS analysis using MQ log(IC_{50}) as the dependent variable and using the mdr1 and plasmepsin 2/3 CNV genotypes as covariates, while controlled for population structure and sample origin. Each point corresponds to a SNP, coloured by the chromosome it is located on and they are ordered by their position on the chromosome. The dotted line indicates the genome-wide threshold for significance (4×10^{-6}). The SNPs exceeding the significance threshold are labeled by the gene they are found in and the amino acid alteration they code for.

Table 3.5: SNPs most strongly associated with MQ log(IC_{50}) when correcting for mdr1 and plasmepsin 2/3 amplifications

Chromosome	Position	Gene ID	Gene Description	N or S	Alteration	p-value
5	961,093	PF3D7_0523000	Multidrug resistance protein 1	N	F1068L	1.9 x10_6
2	438,323	PF3D7_0210800	Conserved Plasmodium protein	N	K227R	1.4 x10_5
2	463,586	PF3D7_0211500	GAF domain-related protein	N	S142G	1.4 x10_5
6	1,068,475	PF3D7_0626400	Sec14 domain containing protein	N	K1110R	4.1 x10_5
14	420,625	PF3D7_1410400	Rhoptry-associated protein 1	N	D62N	7.4 x10_5
10	598,412	PF3D7_1014800	Conserved Plasmodium protein	N	E411A	9.2 x10_5

Table 3.6: SNPs that pass either the Bonferroni-adjusted p-value threshold ($p < 4 \times 10^{-6}$) or the more lenient suggestive threshold ($p < 1 \times 10^{-4}$) are listed in order of increasing p-value. The table shows chromosome and nucleotide position of the SNP, the ID and description of the gene in which the SNP occurs, whether it is a synonymous (S) or nonsynonymous (N) mutation, what amino acid alteration it encodes if it is nonsynonymous, and the p-value associated with the SNP.

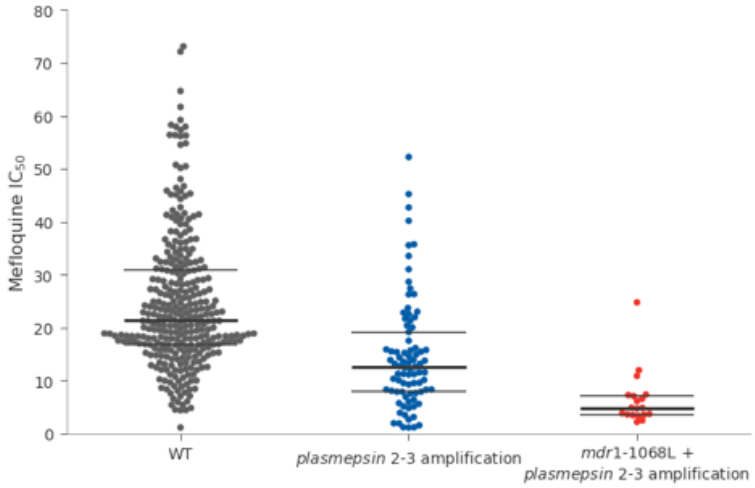


Figure 3.6: Difference in MQ IC_{50} between wild type parasites (grey), those with a plasmepsin 2/3 amplification (blue) and those with both the amplification and the F1068L SNP in the mdr1 gene (red). Samples with an mdr1 amplification have been removed from this analysis. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

to 19% country-wide (figure 3.7). Within Cambodia, the SNP initially appeared in Western Cambodia (WKH) and appears to have spread to Northern Cambodia (NKH) by 2012 (figure 3.7), following the same purported route as the KEL1/PLA1 co-lineage[9,150]. Within NKH, the SNP seems to have increased very rapidly in frequency up to 66% in 2015 (figure 3.7), though the sample number is very low for that year (n = 3). From the available data, the SNP has not yet reached Northeastern Cambodia (NEKH) though no samples were collected in 2015 in this region. Looking at the genomic regions flanking the SNP, I see that they are largely identical (figure 3.8 b) and therefore suggests a single origin of this SNP. I also observe that samples with the SNP are not identical across the genome (figure 3.8 a), indicating that it is not a single clone with this SNP and that the SNP appears to have recombined with a number of different genetic backgrounds. The rapid rise in frequency of the SNP from a single origin suggests that the SNP provides some form of fitness advantage to the parasites harbouring it.

Finally, once I corrected for the F1068L SNP in a final GWAS (figure 3.9 a), I noticed that only

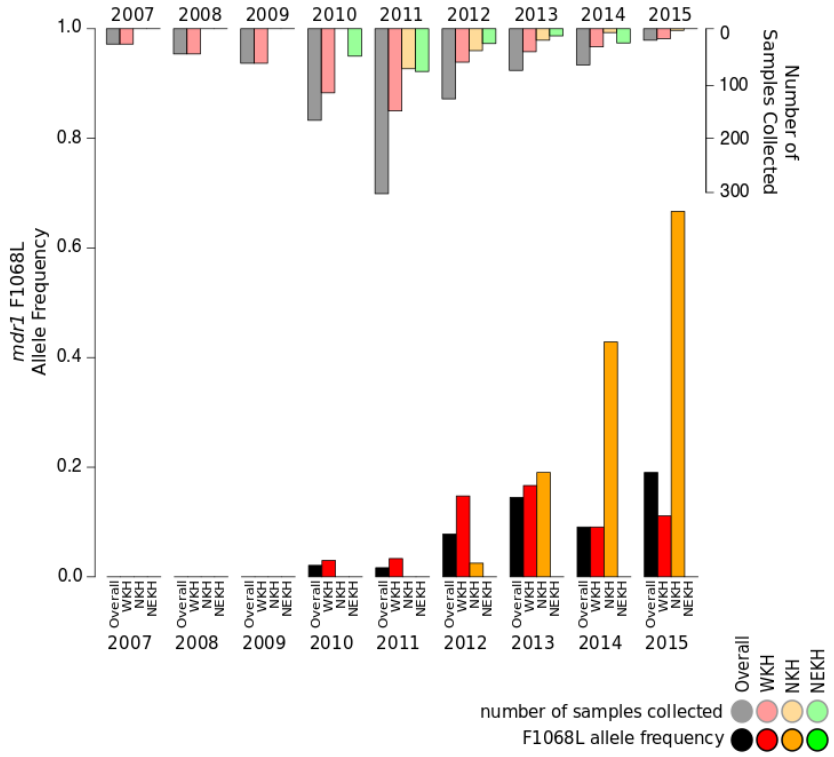


Figure 3.7: Annual allele frequency of the mdr1 F1068L SNP in Cambodia shown in black and by Cambodian region: Western Cambodia (WKH) in red, Northern Cambodia (NKH) in orange, and Northeastern Cambodia (NEKH) in green. The inverted barplot shows the number of samples collected in each region by year.

two SNPs pass the suggestive threshold (table 3.7). These SNPs are located in the same region on chromosome 2, in two genes coding for a conserved protein (K227R SNP in PF3D7_0210800) and a GAF-domain related protein (S142G SNP in PF3D7_0211500). They are in very strong linkage with each other, only 17 samples having differing genotype calls for the two SNPs out of a total of 411 samples with MQ IC_{50} values and both genotypes ascertained. While we see the K227R SNP globally, the S142G SNP is exclusive to Southeast Asia with 80% of samples with the SNP being found in Cambodia (210/264). The SNP is first seen in 2007 at 3% frequency, from where it rapidly increases

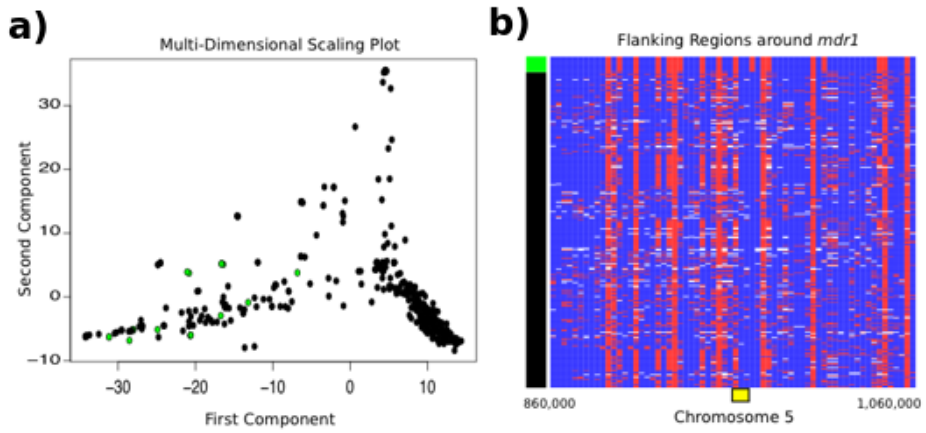


Figure 3.8: a) A multidimensional scaling plot for all 430 samples with MQ IC_{50} values, based on all SNPs called genome-wide. Green dots represent samples with the F1068L mutation. b) Flanking regions around the mdr1 gene, showing the genotype of the 430 samples (rows) for all SNPs within 100kb of either side of mdr1. The genotypes are either reference (blue), alternate (red), or missing (white). The position of the mdr1 gene is indicated with a yellow box. The bar to the left shows which samples have the F1068L mutation in mdr1 (green) and which don't (black).

to roughly 55% in 2015 (figure 3.10). Due to being exclusive to Southeast Asia and the rapid change in frequency, the S142G SNP appears to be the more likely candidate for the observed significant association with MQ IC_{50} in the GWAS and the following analyses were therefore based on the S142G SNP.

The S142G SNP is non-randomly distributed, frequently co-occurring with the *mdr1* F1068L SNP ($p < 5 \times 10^{-5}$), though not necessarily with the *plasmepsin 2/3* amplification ($p > 0.05$) (table

Table 3.7: SNPs most strongly associated with MQ log(IC_{50}) when correcting for mdr1 and plasmepsin 2/3 amplifications as we well as the F1068L mdr1 SNP

Chromosome	Position	Gene ID	Gene Description	N or S	Alteration	p-value
2	438,323	PF3D7_0210800	Conserved Plasmodium protein	N	K227R	1.8×10^{-5}
2	463,586	PF3D7_0211500	GAF domain-related protein	N	S142G	2.9×10^{-5}

Table 3.8: SNPs that pass either the Bonferroni-adjusted p-value threshold ($p < 4 \times 10^{-6}$) or the more lenient suggestive threshold ($p < 1 \times 10^{-4}$) are listed in order of increasing p-value. The table shows chromosome and nucleotide position of the SNP, the ID and description of the gene in which the SNP occurs, whether it is a synonymous (S) or nonsynonymous (N) mutation, what amino acid alteration it encodes if it is nonsynonymous, and the p-value associated with the SNP.

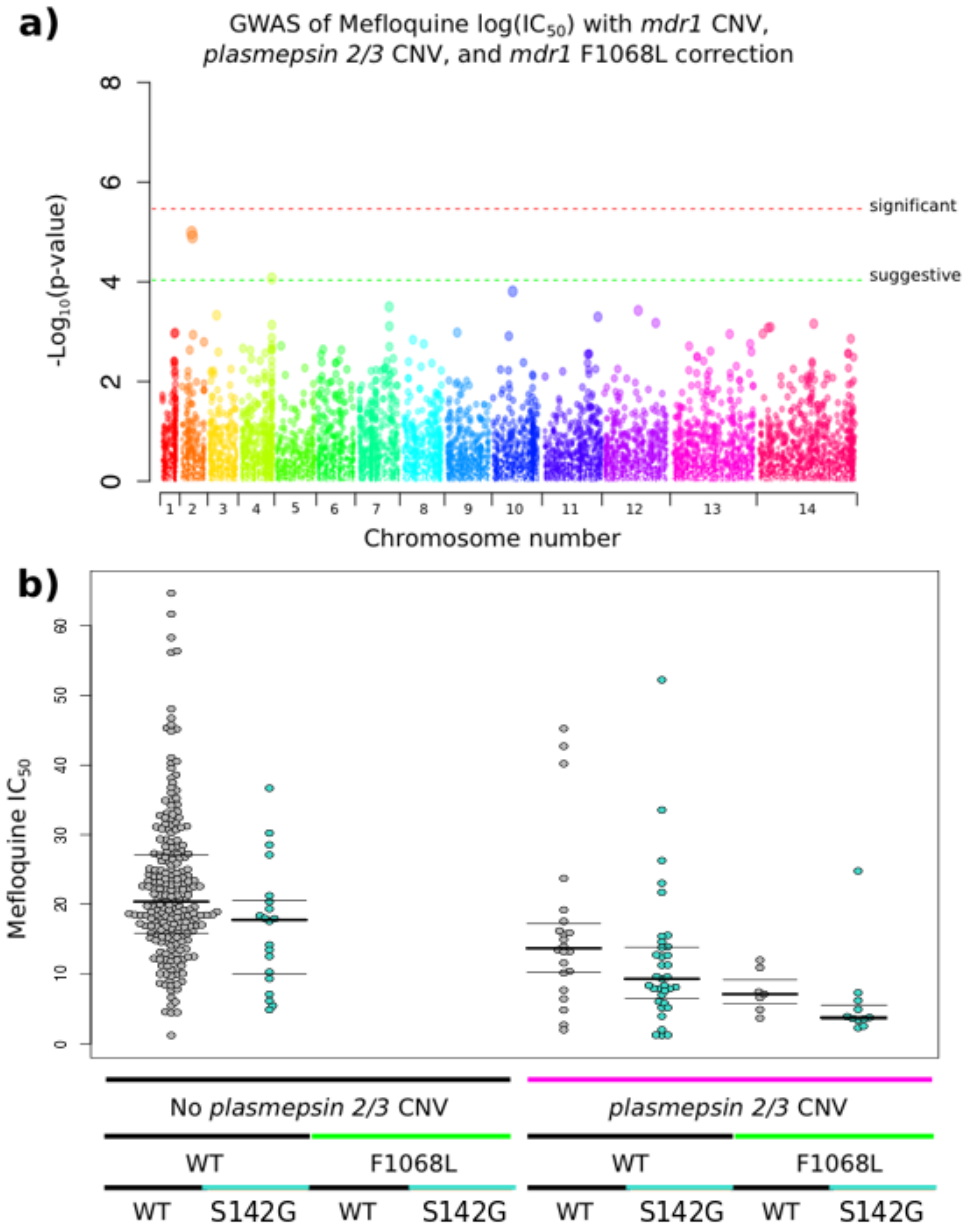


Figure 3.9: a) GWAS analysis using mefloquine log(IC_{50}) as the dependent variable, with population structure, sample origin, mdr1 and plasmepsin 2/3 copy number amplifications, and the F1068L SNP in mdr1 as covariates. Each point corresponds to a SNP, coloured by the chromosome it is located on and they are ordered by their position on the chromosome. The dotted red line indicates the genome-wide threshold for significance (4×10^{-6}) and the dotted green line is a more lenient suggestive threshold (1×10^{-4}). b) Mefloquine IC_{50} values by different combinations of three genetic markers for mefloquine susceptibility: plasmepsin 2/3 amplification (pink), the F1068L SNP in mdr1 (green), and the S142G SNP in PF3D7_0211500 (turquoise). Each point represents a clinical isolate and samples with mdr1 amplifications have been removed. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

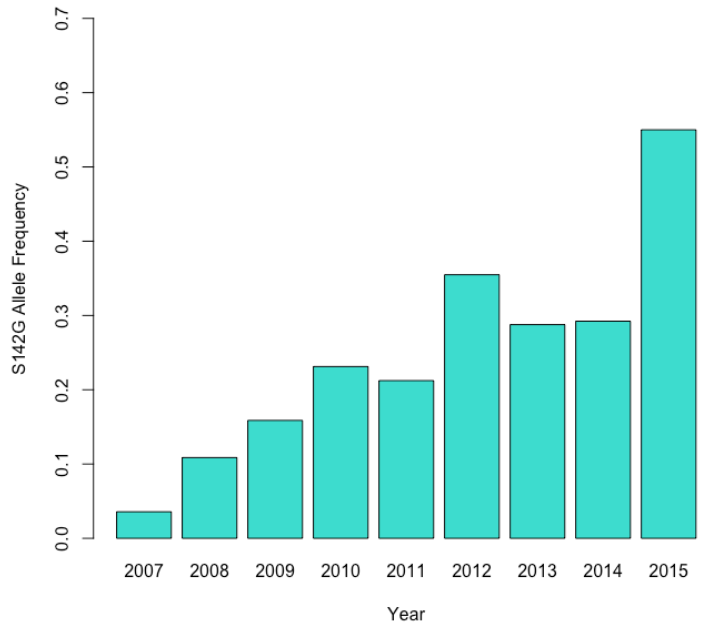


Figure 3.10: Annual allele frequency of the S142G SNP in gene PF3D7_0211500 in Cambodia.

3.9). Additionally, the SNP is very rarely found together with an *mdr1* amplification ($p < 5 \times 10^{-5}$) (table 3.9). In spite of this, the S142G appears to associate with an increase in mefloquine sensitivity in an independent manner to that conferred by the *plasmepsin 2/3* amplification and the F1068L *mdr1* SNP (figure 3.9 b). Looking only at samples without an *mdr1* amplification, I see a significant increase in sensitivity in samples with the S142G SNP compared to those without, in absence of both *plasmepsin 2/3* amplification and the F1068L SNP (16.9 nM ±8.8 vs. 22.3 nM ±10.4, $p < 0.05$), in presence of *plasmepsin 2/3* but absence of F1068L (11.8 nM ±11.9 vs. 16.4 nM ±10.0, $p < 0.05$), and a near significant effect despite the low sample numbers for samples with both the *plasmepsin 2/3* amplification and the F1068L SNP (6.0 nM ±6.4 vs. 7.5 nM ±3.0, $p = 0.056$). This is therefore yet another example of a SNP associating with a particular set of mutations that all appear to result in

sensitizing the parasite to mefloquine treatment.

Table 3.9: Co-occurrence of different markers with the S142G SNP

		S142G SNP	
		Absent	Present
plasmepsin 2/3	Absent	443	134
CNV	Present	174	74
F1068L	Absent	585	177
mdr1 SNP	Present	24	27
mdr1	Absent	442	167
amplification	Present	112	15

Pf6 samples from Cambodia with both genotypes for each comparison ascertained and filtered for QC Pass.
plasmepsin 2/3 CNV: fisher's exact test for count data, $p > 0.05$
mdr1 SNP: fisher's exact test for count data, $p < 5 \times 10^{-5}$
mdr1 amplification: fisher's exact test for count data, $p < 5 \times 10^{-5}$

3.7 Discussion

The current crisis of multidrug resistance in Southeast Asia necessitates an informed, effective, and sustainable response in order to avoid increasing levels of complete treatment failure and, worse, a spread of multidrug resistance to other malaria endemic regions. Complete treatment failure of dihydroartemisinin-piperaquine in Cambodia and other neighbouring countries, caused by the emergence of KEL1/PLA1 parasites, therefore now requires an immediate response. A switch to either artesunate-mefloquine or triple combination therapy, combining artemisinin with both piperaquine and mefloquine, are two options that are currently being explored due to the initial reports of mefloquine efficacy in the field. A number of questions however remained concerning the source of this mefloquine efficacy and the potential sustainability of deploying this drug in a region rampant with multidrug resistance.

In this chapter I have shown that KEL1/PLA1 parasites are not only sensitive to mefloquine, but appear to exhibit a heightened level of susceptibility above and beyond that of wild type parasites,

which we term mefloquine hypersensitivity. I have linked this hypersensitivity to the acquisition of the PLA1 component of the KEL1/PLA1 co-lineage, implying that KEL1/PLA1 are necessarily and inherently hypersensitive to mefloquine. Furthermore, the PLA1 lineage being characterized by the presence of the *plasmepsin 2/3* amplification, I have suggested that the *mdr1* amplification and the *plasmepsin 2/3* amplification may be antagonistic in affecting both mefloquine and piperaquine responses in opposite ways, implying that triple mutants may not necessarily exhibit triple resistance. I have also identified a novel F1068L SNP in the *mdr1* gene that has recently arisen on a *plasmepsin 2/3* background and that has spread through the population, thereby displaying classical hallmarks of positive selection. I have shown that this SNP is associated with an even greater level of mefloquine sensitization than that conferred by PLA1 acquisition, resulting in mefloquine ultrasensitivity. A further novel SNP in a GAF domain-related gene also associates with mefloquine sensitization, complicating the situation even more. The results reported here shed light on the observed efficacy of mefloquine in field, but raise a number of new questions relating to the molecular mechanism of mefloquine sensitization and the evolutionary mechanisms that drive this complex genetic architecture of mefloquine ultrasensitivity.

Mefloquine is thought to act in the cytoplasm by binding to 80S ribosomes and thereby inhibiting protein synthesis[373], with clinical resistance to the drug being mediated by an amplification of the *mdr1* gene[370,70,282], which codes for a transporter protein on the food vacuole membrane. The MDR1 protein purportedly transports mefloquine into the food vacuole and thereby away from its active site, meaning that the higher abundance of MDR1 in *mdr1* amplified parasites would result in more mefloquine being pumped into the food vacuole. This thereby results in reduced mefloquine efficacy and, consequently, clinical resistance to the drug. Similarly, the effect of *mdr1* gene copy

number amplification on increasing piperaquine sensitivity can be explained with piperaquine acting in the food vacuole and the MDR1 protein pumping the drug to where it needs to be. The novel F1068L SNP in *mdr1* that I have identified can similarly be explained by potentially reducing the capability of the MDR1 protein to pump mefloquine into the food vacuole, increasing the effectiveness of the drug. Looking at available data on piperaquine IC_{50} [8], presence of the F1068L SNP does not appear to confer significantly higher levels of piperaquine resistance compared to those without (65.3 nM $\pm$30.6 vs. 58.3 nM $\pm$32.5, $p > 0.05$) (figure 3.11). This suggests that the two drugs are transported in different ways through the MDR1 protein, and that the SNP may therefore be potentially involved in somehow compensating for the *plasmepsin 2/3* amplification with which it co-occurs. The *plasmepsin 2/3* genes code for enzymes present in the food vacuole that are involved in hemoglobin catabolism, specifically conversion of heme to hemozoin [61]. Piperaquine is thought to block this conversion and resistance to the drug is thereby mediated by increasing the availability of the Plasmepsin 2/3 enzymes to overcome the inhibition by piperaquine [372,8]. It is possible that the increased level of Plasmepsin 2/3 enzymes in the food vacuole may require an increased abundance/depletion of other substrates that are imported/exported by the MDR1 protein, a requirement that may be satisfied by a compensatory mutation at the F1068L position. The missing piece in this puzzle of an explanation is the observed effect of *plasmepsin 2/3* amplification on mefloquine sensitivity, as the enzymes are in the food vacuole and the drug acts in the cytoplasm. It is possible that, comparable to the F1068L SNP, the increased abundance of Plasmepsin 2/3 in the food vacuole indirectly affects the type of substrates that the MDR1 protein pumps into the food vacuole. If this change in metabolic activity of the MDR1 protein would result in a reduced level of mefloquine being pumped into the food vacuole, then we would observe a heightened level of mefloquine sensitivity in parasites with a *plasmepsin*

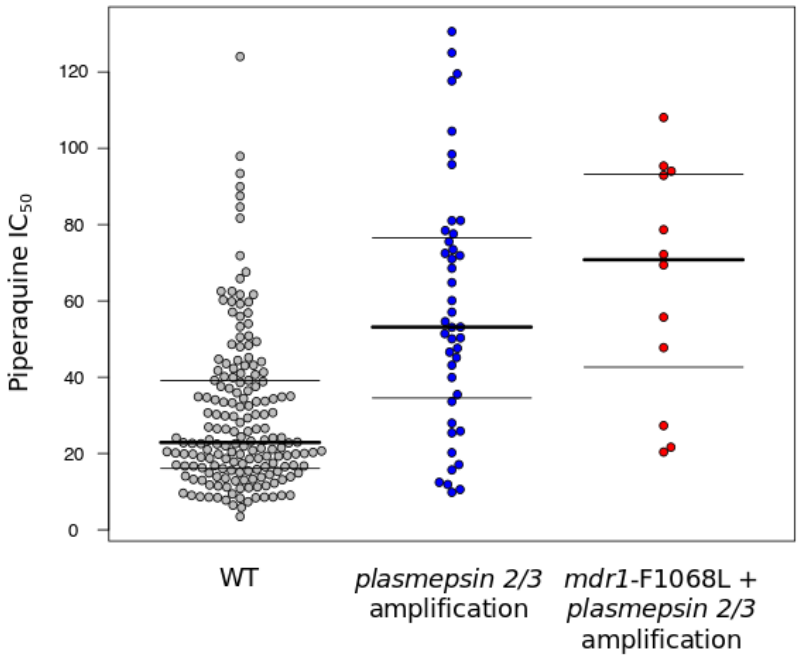


Figure 3.11: Difference in piperaquine IC_{50} between wild type parasites (grey), those with a plasmepsin 2/3 amplification (blue) and those with both the amplification and the F1068L SNP in the mdr1 gene (red). Samples with an mdr1 amplification have been removed from this analysis. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

2/3 amplification.

Thinking about the hyper-sensitization of the parasite population from an evolutionary perspective raises questions about the dynamics of population replacement by hard selective sweeps. The population structure of Southeast Asian *P. falciparum* parasites differs substantially from that of its African counterparts, namely by being significantly more structured. Hard selective sweeps, as driven by selection for chloroquine resistance for instance, have swept through the region in the past. The current outbreak of KEL1/PLA1 is yet another such sweep. The speed at which this spread has occurred is also of significant importance, due to the lag in being able to respond from a policy/logistics perspective. It is especially worrying to see the speed at which KEL1/PLA1 was able to shed the *mdr1*

amplifications that were present in the field at the time[8,9]. It brings up the possibility of the PLA1 component being similarly lost rapidly and replaced with *mdr1* amplification in the case of a policy switch to mefloquine treatment. While the presence of the F1068L *mdr1* SNP may suggest that additional genetic components would have to be dismantled to acquire mefloquine resistance, it is unclear, firstly, how rapidly this could occur and, secondly, whether we wouldn't simply see a new mefloquine resistant lineage combining with the KEL1/non-PLA1 parasites that still circulate in the field. It is also bizarre to see that the KEL1/PLA1 co-lineage arose in West Cambodia, close to the border with Thailand where artesunate/mefloquine was the front line treatment for many years. One would hypothesize that this would have an attenuating effect on parasites that acquire hypersensitivity to mefloquine. It however explains why KEL1/PLA1 parasites did not spread to Thailand in the initial stages of the outbreak. Worryingly, Thailand has now switched the recommended treatment regime to dihydroartemisinin/piperaquine, which would open up the doors to the KEL1/PLA1 co-lineage spreading into the country. Anecdotal reports suggest that the new treatment is less effective in the eastern parts of Thailand than the original treatment (Thanat Chookajorn of Mahidol University Thailand, personal communication). Harnessing the findings of this study should now open up the possibility of better understanding the dynamics that are seen in the field and to better respond by exploiting the underlying interactions between the different drugs that we have at our disposal.

Making the best use of our arsenal of antimalarial drugs is essential for long-term sustainable malaria control and elimination. In order to determine how to most effectively deploy our drugs requires a deep understanding of the interactions between the drugs. In this chapter I have shown that being aware of the antagonistic effect of mefloquine and piperaquine selection pressures on the emergence of the respective drug resistance mutations opens up the door to developing either 'resistance-proof'

drug treatments or to rolling out policy changes that explicitly use this new information to rotate between the drugs for maximum effect and sustainability. In the following chapter I wish to further explore the potential interactions between different antimalarial drugs by analysing available data on four different antimalarial drugs, including chloroquine, mefloquine, piperaquine, and artemisinin.

The Second World War and the conflict in Vietnam brought us most of the drugs available today. The list is small, and the parasite has not been idle: Plasmodium falciparum has now developed resistance to all of our available drugs.

Nicholas White, J. Antimicrob. Chemother., 1992 CE

4

The role of a super-chloroquine resistant *pfcrt* haplogroup in multidrug resistance

4.1 Abstract

Emergence of multidrug resistance in *Plasmodium falciparum* is leading to treatment failure of artemisinin combination therapies in Southeast Asia and is threatening the recent progress made in combating malaria. To unveil novel phenotype-genotype associations and to better understand the parasite pop-

ulation substructuring as a function of multidrug resistance, I harness a large-scale dataset of chloroquine, mefloquine, piperaquine, and artemisinin phenotype data for over 1,000 *P. falciparum* field isolates in conjunction with over 5,000 whole genome sequences from the Pf6 MalariaGEN release (www.malariagen.net). Employing a haplogroup analysis of two well-characterized drug resistance genes, *pfcrt* and *mdr1*, I identify a pair of *pfcrt* haplogroups that exhibit super-resistance to chloroquine. These *pfcrt* haplogroups, characterized by the combination of a CVIET haplotype with the N326S and I356T mutations, also display artemisinin resistance and act as genetic backbones for copy number variations leading to mefloquine and piperaquine resistance. These super-resistant haplogroups are present at high frequency throughout Southeast Asia and may have had a single origin in the western parts of the region. The study highlights the potential for haplogroup analyses to enable us to uncover new biological findings. The findings reported here shed new light on *pfcrt*, one of the most well-studied genes in the *P. falciparum* genome, and potentially suggest that it plays a role in the acquisition of artemisinin resistance.

4.2 Introduction

Antimalarial drug resistance is inevitable. *Plasmodium falciparum*, the apicomplexan parasite that causes the majority of malaria cases worldwide, has developed resistance to every antimalarial drug that it has been confronted with. The parasite mutates on average every base of its genome across the span of a single blood infection[63], an incredible basis for genetic adaptability that has only recently been described. However, the propensity for the parasite to adapt to the challenges imposed by anti-malarial drugs has long been recognized. Beginning with chloroquine in the 1950s[93], *P. falciparum* has subsequently developed resistance to proguanil, sulfadoxine-pyramethamine, mefloquine,

and halofantrine[370]. Each drug lost effectiveness within years of being rolled out, resulting in resistance that still persists to this day as each new generation of drug resistant parasites evolves from the previous one. This is aptly demonstrated by the global pervasiveness of chloroquine resistance currently, even though the drug is officially no longer used to treat *P. falciparum* infections.

With the rollout of artemisinin, the World Health Organization (WHO) recommended switching from mono-therapies to combination therapies. The highly effective but short-lived artemisinin compound was only to be administered in combination with long-lived partner drugs, such as piperaquine. The longer-lived partner drug would clear parasites surviving the artemisinin therapy, preventing resistance from emerging. While logical in theory, *P. falciparum* has again proven its ability to adapt, developing resistance to both artemisinin and to piperaquine[187,330]. In this case, instead of the sequential selection for resistance that occurred with mono-therapies, *P. falciparum* developed resistance concurrently to both drugs. The combination therapy essentially selected for multidrug resistance.

This rise and spread of multidrug resistant malaria parasites has resulted in a paradigm shift in malaria epidemiology. It is no longer sufficient to look at a single drug and a single drug resistance phenotype. Combination therapies are likely here to stay, evidenced by the currently on-going clinical trials of triple-compound therapies[211]. However, the specific combinations of compounds to be used will need to be carefully decided upon in the light of the drug resistance landscape. Traditionally, policy makers have in part relied on epidemiological models to inform them on the best course of action. The resolution and accuracy, and consequently usefulness, of these models depend in large part on our understanding of drug resistance.

We currently have a basic understanding for a variety of key resistance mutations, such as the chloro-

quine resistance transporter (*pfcrt*) K76T mutation for chloroquine resistance [103,82,321], a number of mutations in the *kelch 13* (K13) gene [13,221] for artemisinin resistance, and copy number amplifications of the *plasmepsin 2/3* [8,372] and *multidrug resistance protein 1* (*mdr1*) genes [370,280,282] for piperaquine and mefloquine resistance respectively. While these mutations enable us to differentiate between likely resistant and likely susceptible samples, there is still a significant amount of unexplained variance in the resistance phenotypes. Additionally, how these resistance mutations associate with each other within the complicated multidrug resistance landscape is also little understood. This will be vital in order to uncover antagonistic interactions, such as the *mdr1* amplification (involved in mefloquine resistance) increasing a parasite's susceptibility to piperaquine [8,268] and *plasmepsin 2/3* amplification increasing mefloquine sensitivity as described in Chapter 3.

Increased understanding of factors underlying the variance in drug resistance phenotypes and the interactions between them will be key to making the most informed policy decisions and to designing the most effective combination therapies. By harnessing this fine-grained knowledge, we will be able to turn the parasite's adaptability against itself. Using selective pressures exerted by antimalarial drugs to our advantage will enable us to optimize the use of our existing drugs, and thereby make the treatments 'resistance-proof' [36]. This will be critical, as the antimalarial drug-development pipeline is thin and mostly includes derivatives of existing drugs, to which resistance has often already spread. It is therefore vital that we better understand the drugs currently at our disposal.

In this chapter I describe an analysis of 5,835 whole genome sequences of *P. falciparum* field isolates sampled worldwide in conjunction with previously published phenotype information for four antimalarial drugs: chloroquine, mefloquine, piperaquine, and artemisinin [6,7,15,209]. I apply a haplogroup analysis to classify two known drug resistance genes, *pfcrt* and *mdr1*, and place these within

the larger context of the global drug resistance landscape. I show that two specific *pfcrt* haplogroups exhibit super-resistance to chloroquine and that they share a key N326S mutation in *pfcrt* likely leading to the super-resistance phenotype. Both haplogroups are very common in Southeast Asia, but absent elsewhere in the world, with a gradient of being more common in the western parts of the region than on the eastern side. These super-resistant haplogroups act as genetic backbones for the other types of drug resistance observed in the region: artemisinin, mefloquine and piperaquine resistance. I discuss the implications of these findings and suggest a possible role for *pfcrt* in the process of developing artemisinin resistance in the Southeast Asian setting.

All methods used for this chapter are detailed in Appendix F. All sample collection, sequencing, and *in vitro* drug resistance assays were performed by collaborators. All data analyses presented in this chapter are my own work unless specifically noted otherwise.

4.3 Data Overview

To better understand the genetic basis of multidrug resistance in Southeast Asia, as well as to place the findings within a global context, I performed a comprehensive haplogroup analysis of two well-studied drug resistance genes, *pfcrt* and *mdr1*, using the globally sampled Pf6 dataset (www.malariagen.net). These two genes are attractive targets to use for this study for a number of reasons. Firstly, their functional roles in resistance to chloroquine [103,82,321] and mefloquine [370,280,282], respectively, are well-studied and thereby give us a solid basis to place any results into context with previous findings. Secondly, both genes have previously been implicated in studies of other antimalarial drugs, such as artemisinin for *pfcrt* [221] and piperaquine for *mdr1* [8]. Finally, both *pfcrt* and *mdr1* genes code for transmembrane proteins located on the membrane of the digestive vacuole, mediating intake and efflux of

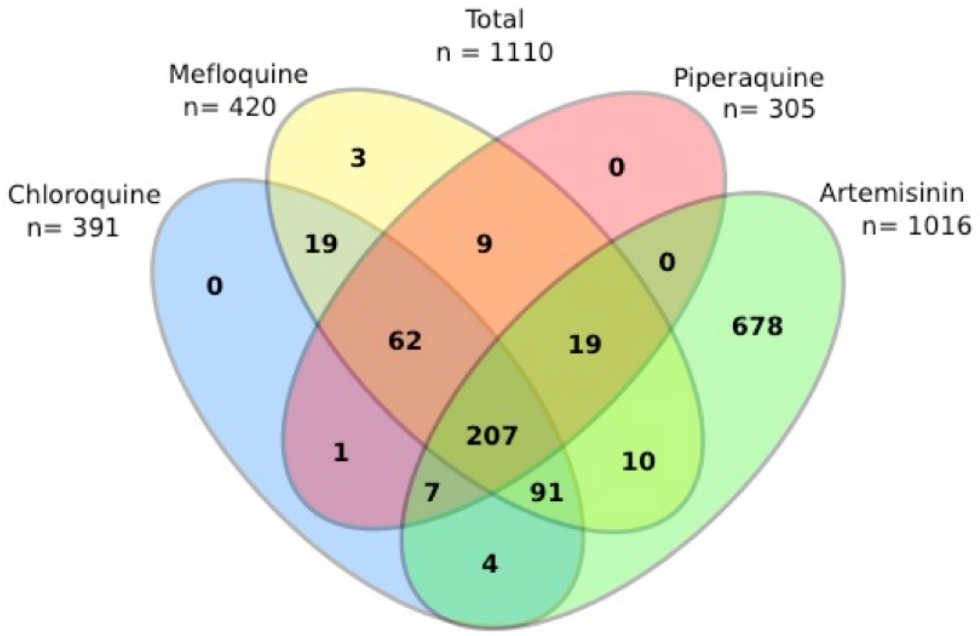


Figure 4.1: Distribution of phenotype data among the 1110 clinical P. falciparum isolates with at least one antimalarial resistance phenotype assayed. Numbers below antimalarial compound indicate number of samples assayed for that compound. Of the 1110 samples, 207 have been assayed for all four compounds.

compounds, and are thought therefore potentially to interact on a mechanistic level through shared biological pathways. A better understanding of the impact of different haplotypes of *pfcrt* and *mdr1*, as well as the specific pairing of these, on different drug resistance phenotypes is therefore needed.

The ongoing sequencing projects led by the MalariaGEN consortium have led to the assembly of a dataset consisting of 5,835 *P. falciparum* whole genome sequences (WGS) (www.malariagen.net). A number of these samples were collected as part of different drug resistance surveillance studies[6,7,15,209] and as a result have drug resistance phenotype information associated with them in the form of 50% inhibitory concentration (IC_{50}) values for chloroquine (CQ) (n = 391), mefloquine (MQ) (n = 420), and piperaquine (PPQ) (n = 305), and in the form of parasite clearance half-life ($PC_{t1/2}$) values for artemisinin (ART) (n = 1,016) (figure 4.1). In total, 1,110 *P. falciparum* isolates have at least one antimalarial drug resistance phenotype assayed and 207 have phenotypes assayed for all four antimalarial compounds (figure 4.1).

To perform the haplogroup analysis, Jacob Almagro Garcia from the Big Data Institute in Oxford classified samples by all their nonsynonymous mutations, as well as insertions/deletions, for *pfcrt* and

mdr1 independently (Appendix F). In terms of terminology, I employ the term 'haplogroup' here instead of the more conventional 'haplotype' because the ultimate aim of this particular study is to group similar haplotypes into 'haplogroups' that in themselves likely represent a single evolutionary origin. While at the point of writing up this chapter the grouping of haplotypes has not yet been undertaken, I have nonetheless chosen to proceed with the final terminology of 'haplogroup'. To increase the robustness and reliability of the classification, only WGS with no missing genotype calls within the target genes were included (4,424 and 4,573 samples for *mdr1* and *pfcrt* respectively). Heterozygous calls in the respective genes (indicative of mixed genotype infections) were set to missing and thus removed from being classified. It is important to note that it is possible that in cases of a mixed genotype infection with a dominant strain, either or both of the genes could be classified if it happens that none of the SNP calls in the genes are called as heterozygous. Each sample was thereby assigned to a haplogroup for that particular gene, within which every sample has an identical amino acid sequence, coded for by an identical set of mutations. Using this method, 113 unique *mdr1* haplogroups (figure 4.2) and 49 unique *pfcrt* haplogroups (figure 4.3) were identified.

4.4 Description of the *mdr1* Haplogroups

The *mdr1* haplogroups have on average two genetic differences to the *mdr1* 3D7 reference sequence (figure 4.2). Only a single genetic change was present in over half of *mdr1* haplogroups (65/113), namely the Y184F mutation. Other mutations common to a number of different haplogroups include N86Y (15/113), N1042D (8/113), and D1249Y (8/113). However, the majority of the genetic changes differentiating the different *mdr1* haplogroups were exclusive to specific haplogroups (42/70). The two most frequent *mdr1* haplogroups (*mdr1*.h1 and *mdr1*.h2) are present in over half

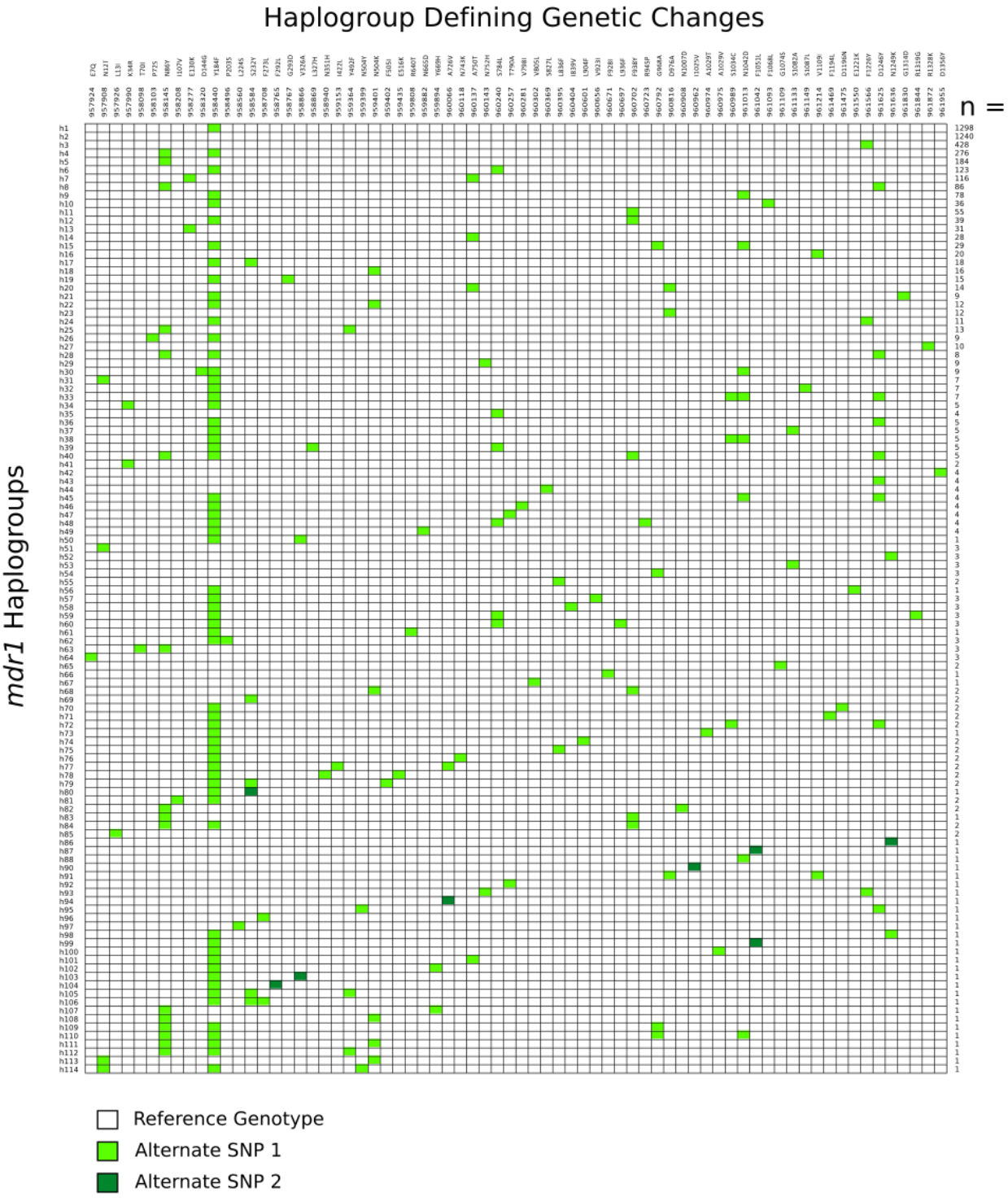


Figure 4.2: All 113 mdr1 haplogroups with the genetic changes that characterize them. Columns indicate genetic changes, labeled with the nucleotide position of the change and the primary amino acid change (multi-allelic positions may result in an alternate change). Colours represent the genetic changes, with white representing no change, light green representing a SNP and dark green representing an alternate SNP in a multi-allelic position. The number of samples in each haplogroup is indicated on the right side of the diagram.

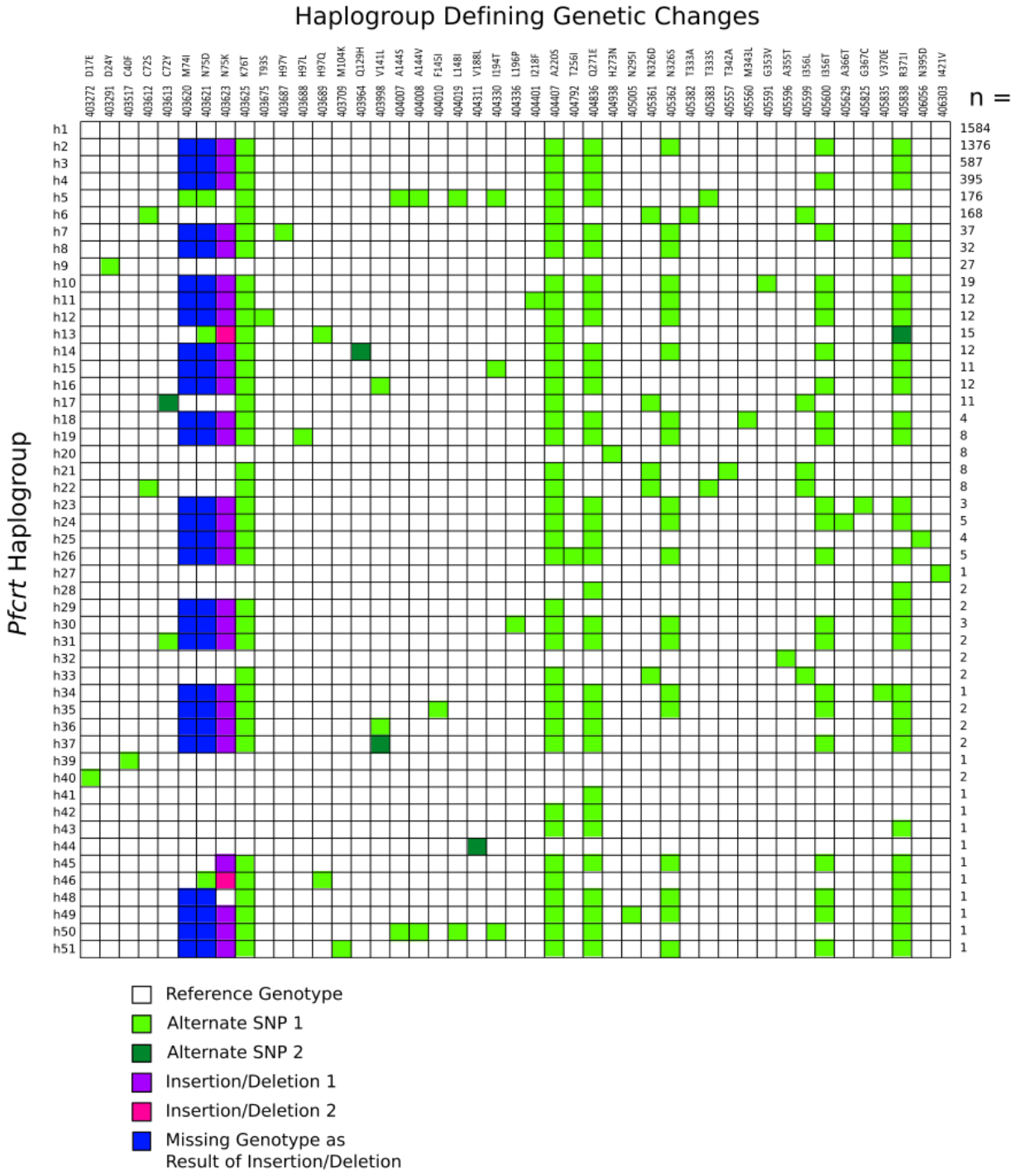


Figure 4.3: All 49 pfcrt haplogroups with the genetic changes that characterize them. Columns indicate genetic changes, labeled with the nucleotide position of the change and the primary amino acid change (multi-allelic positions may result in an alternate change). Colours represent the genetic changes, with white representing no change, light green representing a SNP and dark green representing an alternate SNP in a multi-allelic position. The purple and pink represent different insertions/deletions, while blue indicates positions coded for as missing due to an insertion/deletion. The number of samples in each haplogroup is indicated on the right side of the diagram.

of the samples (2,538/4,424) (figure 4.4 a), with the latter being identical to the 3D7 reference *mdr1* sequence and *mdr1*.h1 differing from the reference only by the Y184F mutation. They are both found throughout Africa and Southeast Asia and of the 113 *mdr1* haplogroups, only 29 are found in more than one *P. falciparum* population, the others being region specific (figure 4.4 b). A large proportion of these region-specific *mdr1* haplogroups are found in West Africa (35/84), however these tend to be haplogroups with very few samples. I noticed that a number of larger haplogroups are specific to Eastern Southeast Asia, including *mdr1*.h10 (36 samples), *mdr1*.h16 (10 samples), *mdr1*.h19 (15 samples), *mdr1*.h20 (14 samples), and *mdr1*.h21 (9 samples) among others (figure 4.4 a & b).

Focusing on the ten largest *mdr1* haplogroups, which account for over 85% of samples (3865/4424), I see that they are generally closely related to each other (figure 4.5 a) and that only nine mutations differentiate all of them (figure 4.5 c). Looking at the MQ IC_{50} by haplogroup, I don't see any significant difference between the two largest *mdr1* haplogroups (t-test: $p > 0.05$), but I observe that *mdr1*.h10 has a significantly lower mean MQ IC_{50} compared to them ($p < 0.0005$) (figure 4.5 b). This particular haplogroup is exclusive to Eastern Southeast Asia (figure 4.5 d) and has a mutation that is specific to this haplogroup (F1068L) and was described in the previous chapter in a genome wide association analysis of mefloquine resistance using the same dataset. As described there, the F1068L mutation seems to have recently emerged and is rising in frequency. The first samples with the mutation are from 2010, where they accounted for about 1% of the total samples from Eastern Southeast Asia (3/276), by 2012 it was 5% (10/189), and by 2015 almost 20% of samples carried this mutation (4/21).

Looking at the distribution of the phenotype data for the three other antimalarial drugs, I see large differences in the average level of artemisinin resistance between the different *mdr1* haplogroups (fig-

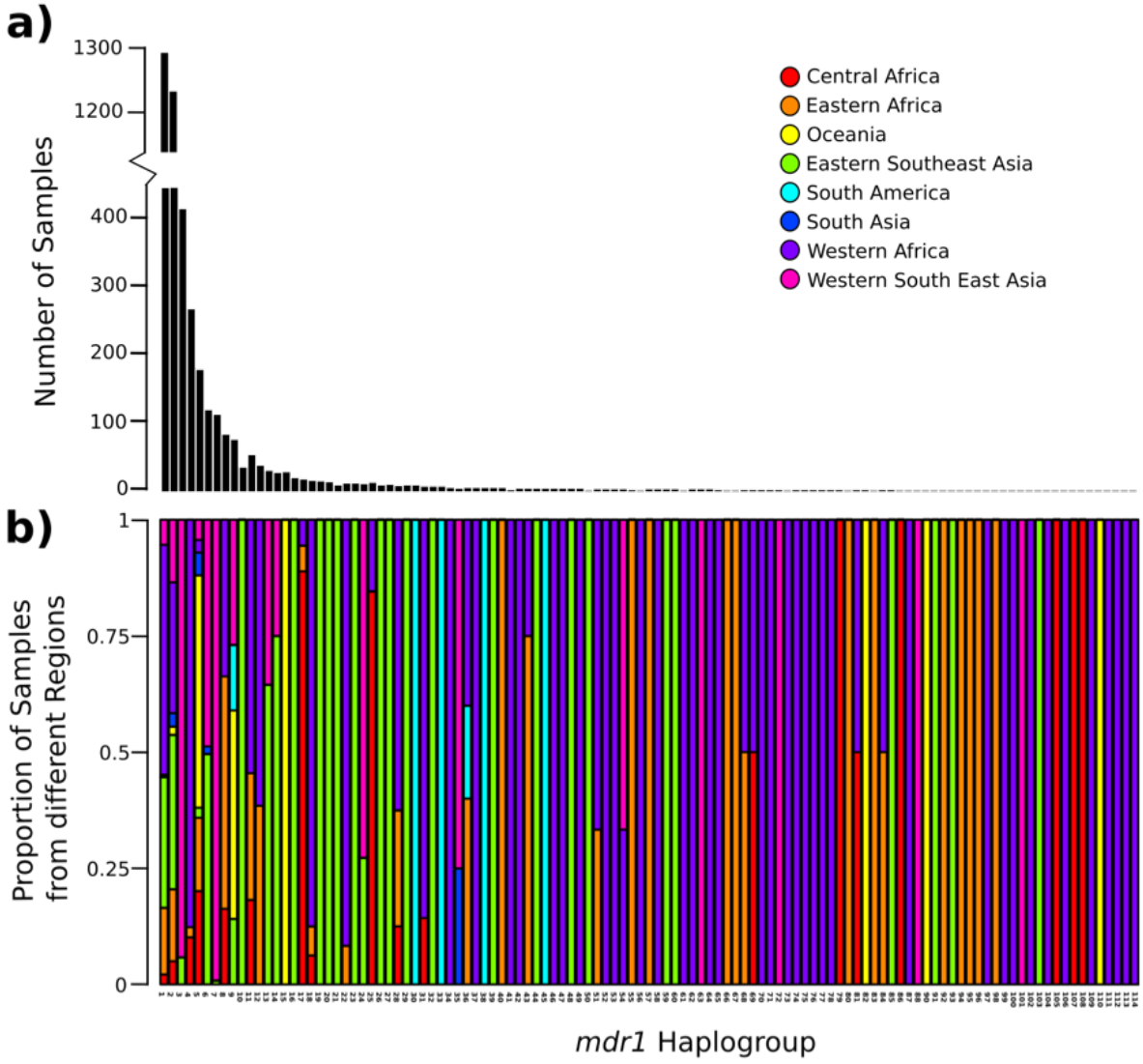


Figure 4.4: Showing the number of samples for each mdr1 haplogroup (a) and the proportion of samples from the different P. falciparum populations for each mdr1 haplogroup (b).

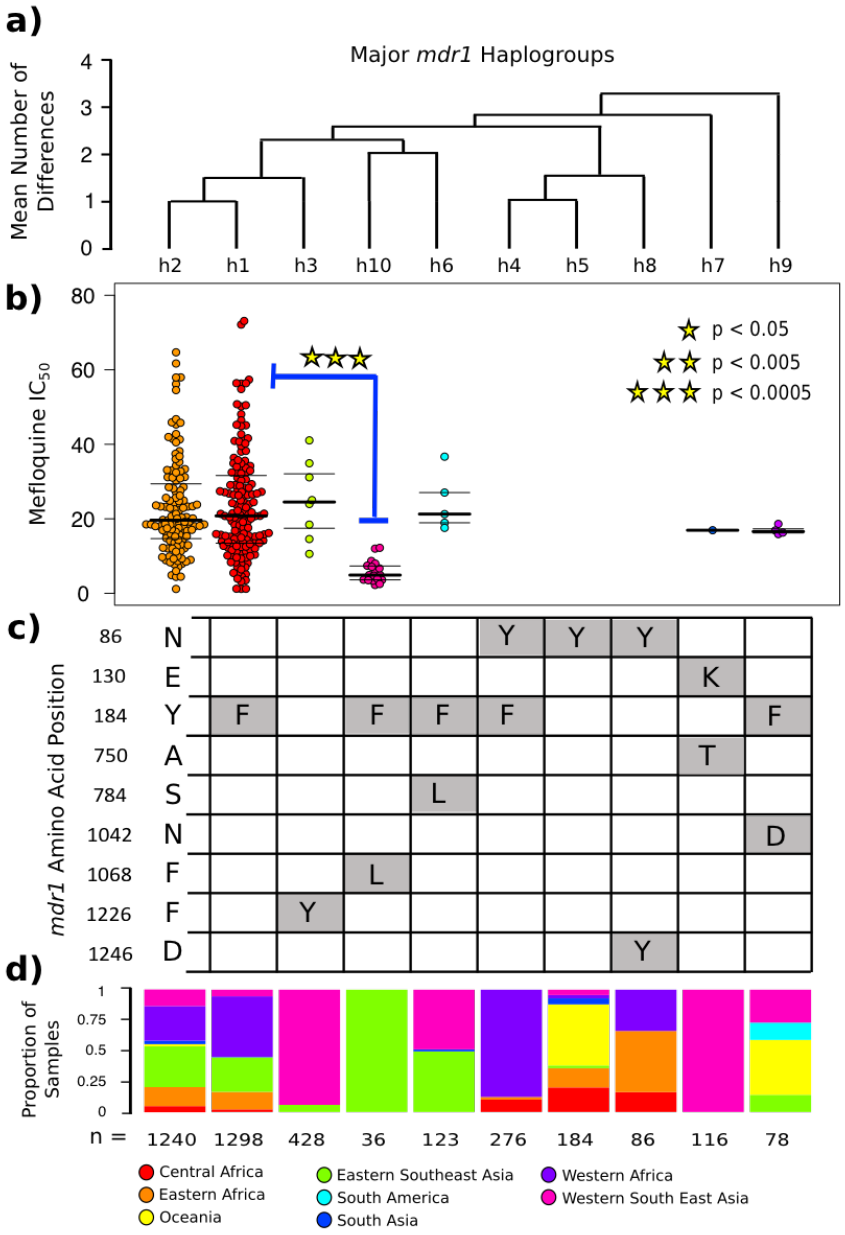


Figure 4.5: Features of the major mdr1 haplogroups. a) Tree showing the genetic similarity of the different mdr1 haplogroups to each other, with branch height indicating the mean number of genetic differences between nodes. b) Mefloquine IC_{50} by mdr1 haplogroup. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (see inset legend). c) The genetic differences between the different mdr1 haplogroups, with rows indicating different amino acid positions within the mdr1 protein. The first column (mdr1.h2) shows the reference amino acid sequence. Empty cells correspond to the reference amino acid as well, while grey boxes show the new amino acid when there was a change. d) Bars indicate the proportion of samples originating from the different P. falciparum populations (see legend) for each mdr1 haplogroup. The numbers below the bars indicate the number of samples in each of the haplogroups.

ure 4.6 a). Specifically, *mdr1*.h4, *mdr1*.h5, and *mdr1*.h8 are all significantly more susceptible to artemisinin than *mdr1*.h1 (all t-test: $p < 0.005$) (figure 4.6 a). These three haplogroups are related to each other by sharing the N86Y mutation (figure 4.5 c), and indeed there is a very significant difference in the level of artemisinin resistance between samples with the wildtype N and the mutant Y genotype (3.9hr vs. 2.2hr, $p < 5 \times 10^{-16}$). While this is confounded by the fact that the haplogroups containing the mutant genotype are not found in Southeast Asia (figure 4.5 d) and therefore appear to lack the *kelch 13* mutations, I still find a significant difference between the two groups when I exclude samples with *kelch 13* mutations (3.0hr vs. 2.2hr, $p < 1 \times 10^{-5}$) (figure 4.6 b). However, we observe that this difference is driven by samples from Eastern Southeast Asia and that when I remove all samples from the Southeast Asian region, that the difference in artemisinin resistance disappears (2.3hr vs. 2.2hr, $p > 0.05$). This suggests that there are additional factors that drive artemisinin resistance within Southeast Asia other than the *kelch 13* mutations, but that these are apparently unrelated to the *mdr1* haplotype.

Finally, I also observe significant differences between *mdr1* haplogroups in their average level of piperaquine IC_{50} (figure 4.6 c) and chloroquine IC_{50} (figure 4.7 a). For the former, *mdr1*.h1 has a significantly higher level of piperaquine IC_{50} than *mdr1*.h2 (41nmol/L vs. 31nmol/L, $p < 0.01$) and *mdr1*.h3 (41nmol/L vs. 22nmol/L, $p < 0.005$) (figure 4.6 c). This is likely due to the overrepresentation of the *plasmepsin 2/3* amplification in this haplogroup (10%, 131 out of 1,279) compared to the other two (both <1%, 8 out of 1,230 and 1 out of 424 for *mdr1*.h2 and *mdr1*.h3 respectively). For chloroquine IC_{50}, I observe a significant difference between *mdr1*.h2 with *mdr1*.h1 and *mdr1*.h10 (figure 4.7 a), which differ from each other by the Y184F mutation (figure 4.5 c). While I do see a significant difference between samples with the Y184F mutation compared to those that

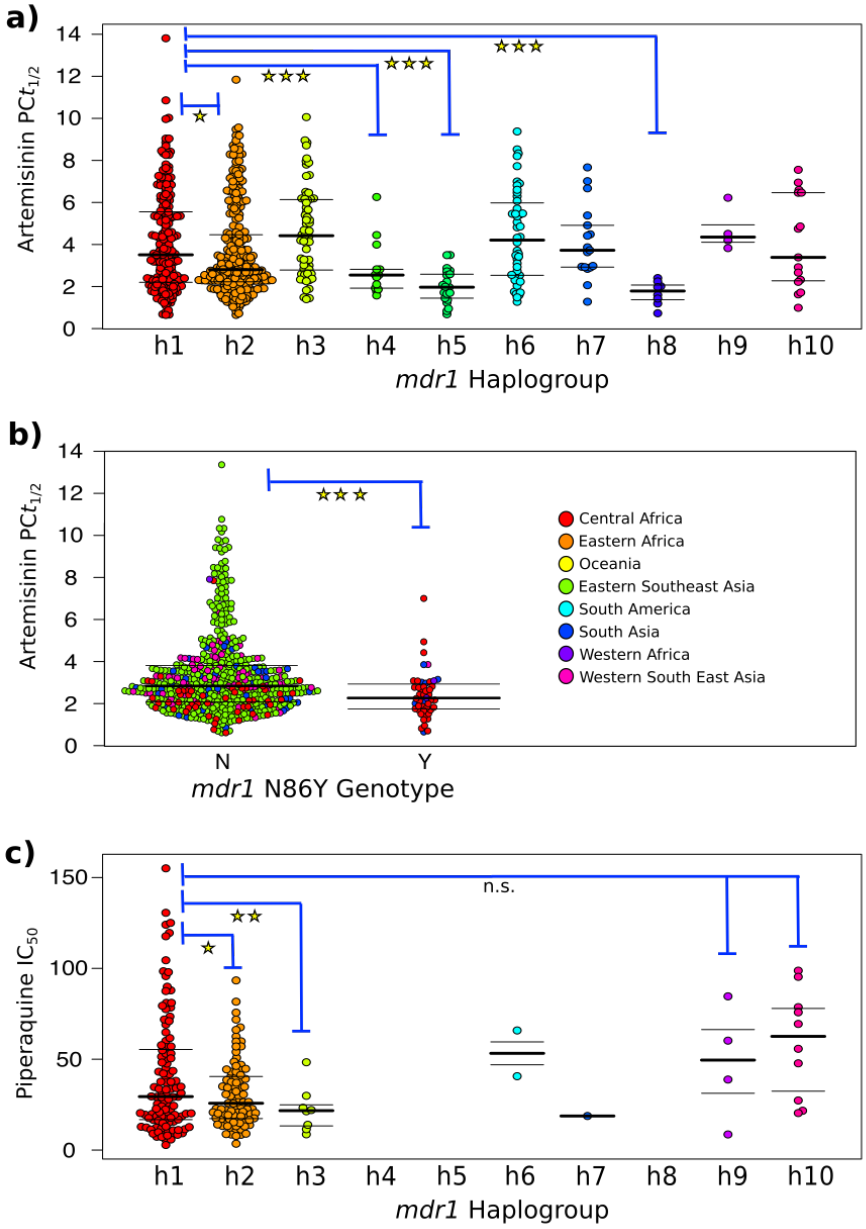


Figure 4.6: Artemisinin $PC_{t1/2}$ by mdr1 haplogroup (a) and by mdr1 N86Y genotype for samples without kelch 13 mutation (b) and piperaquine IC_{50} by mdr1 haplogroup (c). The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (n.s. = p > 0.05, * = p < 0.05, ** = p < 0.005, *** = p < 0.0005). In (b), the colors of the samples correspond to the P. falciparum population they originate from (see inset legend).

don't (441nmol/L vs. 277nmol/L, $p < 5 \times 10^{-13}$) (figure 4.7 b), this difference is confounded by the differential distribution of *pfcrt* haplogroups (see below) (figure 4.7 b). When I stratify by *pfcrt* haplogroup I do not observe a difference in chloroquine IC_{50} between *mdr1* haplogroups (figure 4.7 c). These results suggest that *mdr1* haplogroup may have little effect on the level of resistance to antimalarial drugs other than mefloquine and shows to which extent the population structure can act as a confounder.

4.5 Description of the *pfcrt* Haplogroups

With an average of six genetic differences, the 49 identified *pfcrt* haplogroups are overall significantly more diverged from the *pfcrt* reference sequence than the *mdr1* haplogroups were from the *mdr1* reference sequence (t-test: $p < 0.01$) (figure 4.3). The largest haplogroup (*pfcrt*.h1) consists of 1,584 samples and is identical to the 3D7 reference strain, while 13 of the 49 haplogroups contain only a single sample (figure 4.8 a). The eight largest haplogroups each contain over 30 samples and collectively account for over 95% of all samples (4,355/4,573) (figure 4.8 a). Of the 49 haplogroups, 37 contain the K76T mutation and of these, 27 have a CVIET haplotype. Interestingly, the CVIDT haplotype is found in only one haplogroup (*pfcrt*.h5) and contains a number of ancillary mutations (A144S, A144V, L148I, I194T, T333S) that distinguish it clearly from all other haplogroups (figure 4.9 a), suggesting that the CVIDT haplotype may have had a single origin. The only exception to this observation is *pfcrt*.h50 (containing only one sample), which has a CVIET haplotype but does contain a number of these ancillary mutations and may be a recombinant of CVIDT and CVIET parents. Of the 27 CVIET haplogroups, 25 contain the A220S, Q271E, and R371I mutations, with the remaining two haplogroups either lacking R371I (*pfcrt*.h25) or Q271E (*pfcrt*.h30). There is therefore strong

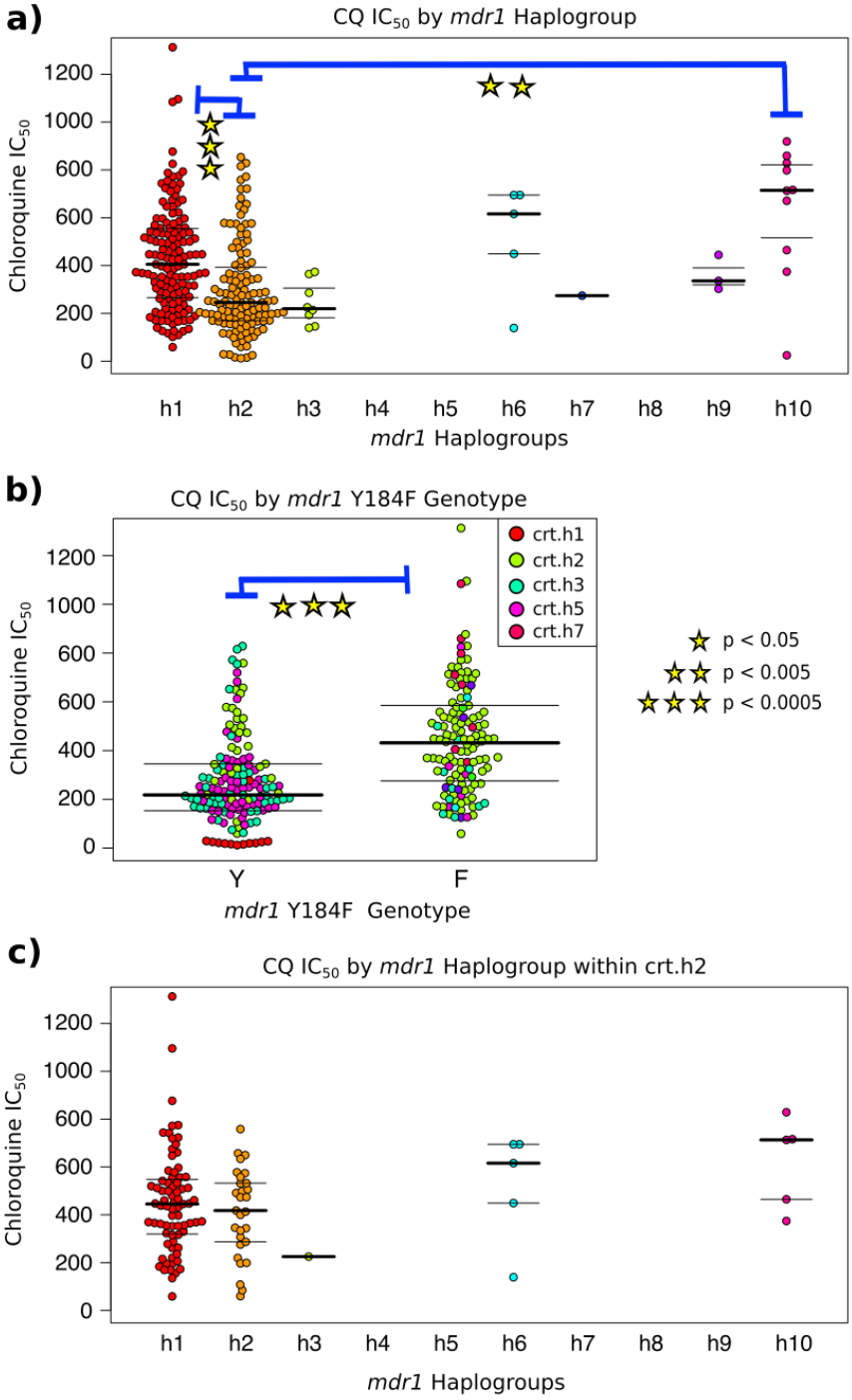


Figure 4.7: Chloroquine IC_{50} by mdr1 haplogroup (a), by mdr1 Y184F genotype (b) and by mdr1 haplogroup only within samples with a pfcrt.h2 haplogroup (c). The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (n.s. = $p > 0.05$, * = $p < 0.05$, ** = $p < 0.005$, *** = $p < 0.0005$).

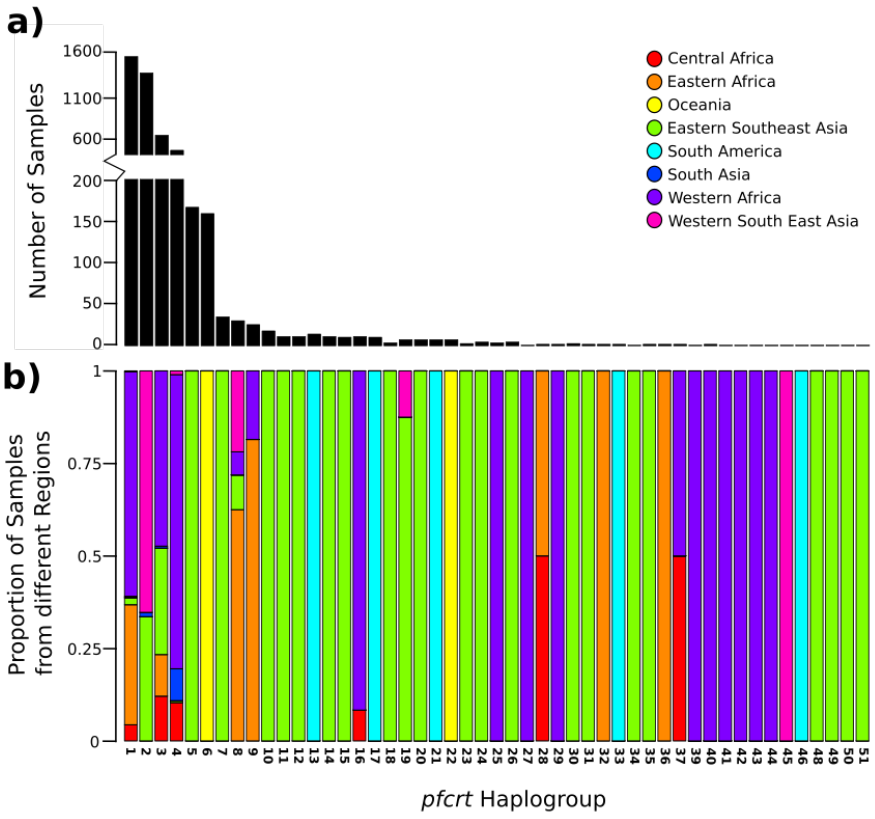


Figure 4.8: Showing the number of samples for each pfcrt haplogroup (a) and the proportion of samples from the different P. falciparum populations for each pfcrt haplogroup (b).

linkage of the CVIET haplotype with these accessory mutations and it will be difficult to disentangle the respective contributions of these mutations to the resistance phenotype.

Two additional mutations are common to a number of CVIET haplogroups (N326S and I356T), with 18 CVIET haplogroups carrying N326S and 20 CVIET haplogroups carrying I356T, and 17/27 carrying both. Over half of all genetic changes differentiating the different haplogroups (26/45) are single mutations unique to specific haplogroups. Additionally, looking at the structure of the *pfcrt* protein, a large number of the genetic changes are located in transmembrane domains (34/45) (figure 4.9 b), which is a significant enrichment compared to the overall fraction of the protein that is transmembrane (227/424) (chi-squared test: $p < 0.01$). This was even more evident when looking only at the genetic changes that are found in multiple haplogroups (17/19) (chi-squared test: $p < 0.005$). The genetic changes mainly occur in the first three and in the eighth and the ninth transmembrane

domains (out of a total of ten transmembrane domains) (figure 4.9 b).

Looking at the geographical distribution of the *pfcrt* haplogroups using information on the sample origin revealed a number of interesting patterns. Firstly, the wild type *pfcrt*.h1 is common in Africa, being the most common haplogroup in a number of African countries, but is underrepresented in most other parts of the world (figure 4.10). The most widespread haplogroup besides *pfcrt*.h1 seems to be *pfcrt*.h3, which has a high frequency in both Africa and parts of Asia, and is the quintessential CVIET haplogroup, containing the CVIET haplotype and the three ancillary mutations. No haplogroup with a CVIET haplotype is observed in South America or in Oceania. I also observe that regions outside Africa tend to be dominated by haplogroups specific to those regions, such as *pfcrt*.h2 in Southeast Asia, *pfcrt*.h6 in Oceania, and *pfcrt*.h13 and *pfcrt*.h17 in Colombia and Peru respectively (figure 4.10). Within Africa, Ethiopia differs from the other African countries in having the *pfcrt*.h8 haplogroup as the most common *pfcrt* haplogroup, despite being virtually absent elsewhere. It is interesting to note that while I saw striking differences in the *pfcrt* haplogroups between the different regions and even between different countries, little diversity in haplogroups was found within countries, most places only harbouring a handful of haplogroups. The exception to this was Eastern Southeast Asia, which exhibited a large number of different haplogroups, including the CVIDT haplogroup, which is exclusive to that region (figures 4.8 b & 4.10).

4.6 Overlap of *mdr1* and *pfcrt* Haplogroups

Looking at the overlap of the *mdr1* haplogroups with the *pfcrt* haplogroups, I notice a number of interesting features. Focusing on haplogroups that have at least ten samples with both *mdr1* and *pfcrt* haplogroups confidently called, it appears that the two sets of haplogroups do not pair up randomly

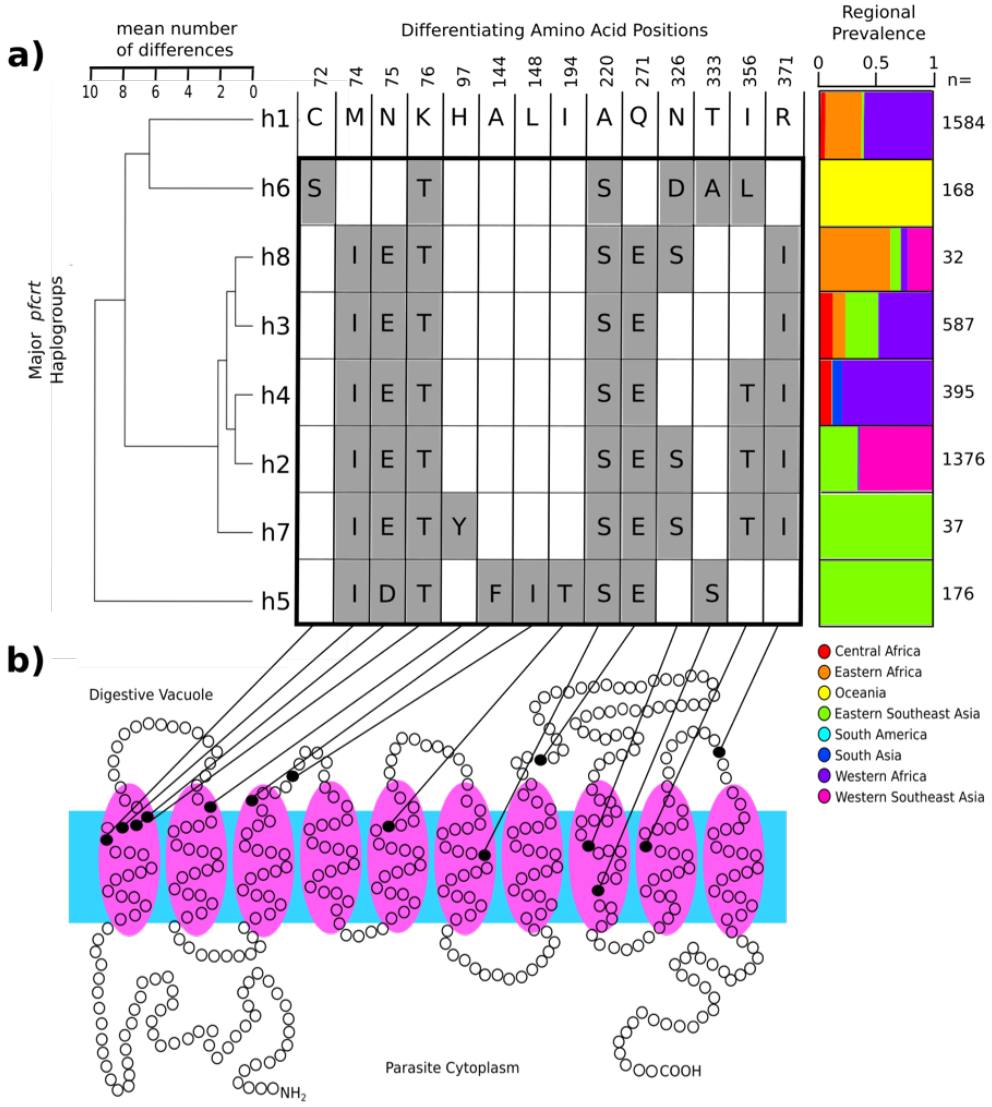


Figure 4.9: Features of the major pfcrt haplogroups. a) Left: Tree showing the genetic similarity of the different pfcrt haplogroups to each other, with branch height indicating the mean number of genetic differences between nodes. Center: The genetic differences between the different pfcrt haplogroups, with columns indicating different amino acid positions within the pfcrt protein. The first row (pfcrt.h1) shows the reference amino acid sequence. Empty cells correspond to the reference amino acid as well, while grey boxes show the new amino acid when there was a change. Right: Bars indicate the proportion of samples originating from the different P. falciparum populations (see legend below) for each pfcrt haplogroup. The numbers on the right indicate the number of samples in each of the haplogroups. b) Showing a model structure of the PfCRT protein, with its ten transmembrane domains (pink) sitting within the membrane of the digestive vacuole (blue). Each circle corresponds to an amino acid within the protein, with shaded circles referring to amino acids that vary between the eight major pfcrt haplogroups, with a line linking these back to the previous panel.

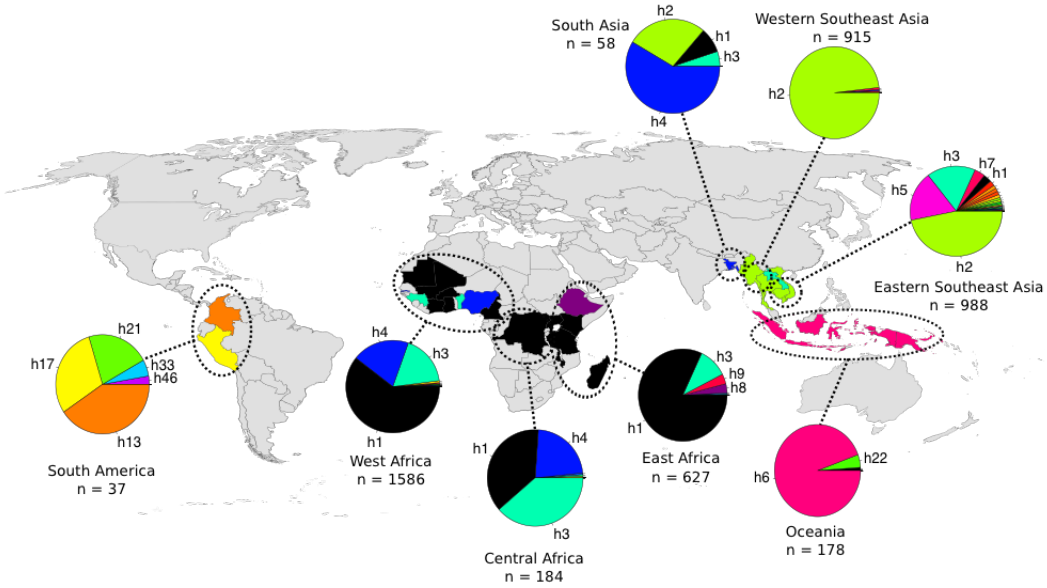


Figure 4.10: World map showing the distribution of the different pfcrt haplogroups. Countries are coloured by the most prevalent pfcrt haplogroup in the country. Dashed circles indicate different P. falciparum populations, with pie charts showing the proportions of the different pfcrt haplogroups within those populations. The number of samples in each of those populations is indicted next to the pie chart.

(chi-squared test: $p < 10^{-16}$). I find that specific combinations of haplogroups are either strongly over-represented or strongly underrepresented (hypergeometric test: $p < 10^{-15}$) (figure 4.11). For *pfcrt*.h1, which is the reference sequence *pfcrt*, it pairs up more often than expected with *mdr1*.h1, but almost never pairs up with *mdr1*.h3, *mdr1*.h6, *mdr1*.h7, and *mdr1*.h9. This is as expected from their respective distributions, as *pfcrt*.h1 is rare outside Africa and the latter four *mdr1* haplogroups are not found in Africa. The opposite pattern is found for *pfcrt*.h2, which is very prevalent in Southeast Asia and pairs very often with *mdr1*.h3, *mdr1*.h6 and *mdr1*.h7. Indeed, many of the signals of association for the pairs of haplogroups are driven by the differential geographical distribution of the haplogroups.

Some of the more unexpected combinations are exhibited by the smaller haplogroups. I found that *mdr1*.h19 only occurs in conjunction with *pfcrt*.h2, which may be a result of a clonal expansion related to the *plasmepsin 2/3* copy number amplification present in all these samples (figure 4.11). The same may have occurred for *pfcrt*.h10, which was only found together with *mdr1*.h1, and with *pfcrt*.h12 only co-occurring with *mdr1*.h10, all of which have the *plasmepsin 2/3* amplification. It is therefore interesting to note that it was possible to use haplogroup information to dissect the samples to get at

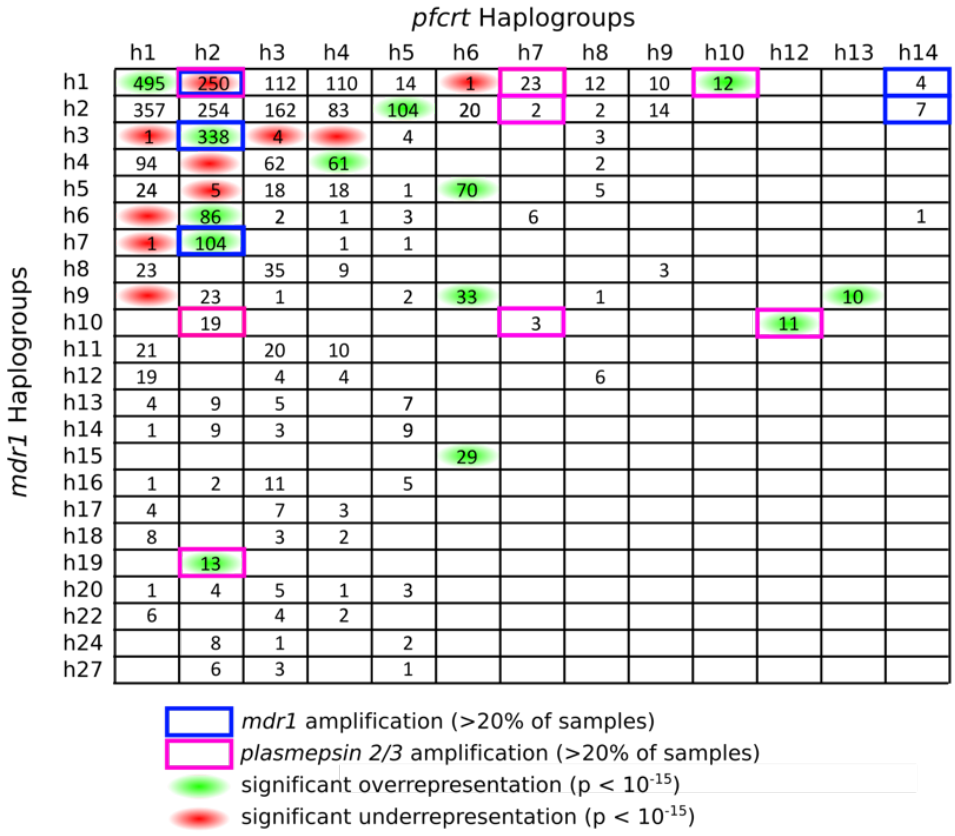


Figure 4.11: Number of samples containing certain combinations of pfcrt and mdr1 haplogroups. Empty cells indicate that no sample had that specific combination of haplogroups. Cells that have over 20% of samples with mdr1 copy number amplifications are indicated with blue squares, while those with over 20% of samples exhibiting plasmepsin 2/3 copy number amplifications are indicated with pink squares. Haplogroup combinations that are significantly overrepresented or underrepresented are indicated with green and red backgrounds respectively (hypergeometric test: $p < 10^{-15}$).

observations relating to their geographical origins and their potential drug resistance phenotypes.

4.7 A Super Chloroquine Resistant Haplogroup

Using the available phenotype data, I investigated whether the eight major *pfcrt* haplogroups differ in their susceptibility to chloroquine (figure 4.12 a). Of the five major *pfcrt* haplogroups for which I had CQ IC_{50} values, I found that the samples with the *pfcrt*.h1 haplogroup were as expected very susceptible to chloroquine due to carrying the wild type *pfcrt* haplotype (42nmol/L vs 373nmol/L, $p < 4 \times 10^{-11}$). I also found significant differences in CQ IC_{50} values between the haplogroups carrying the CVIET haplotype (figure 4.12 a). All three haplogroups are chloroquine resistant, however they differ in their levels of resistance. The archetypical CVIET haplogroup, *pfcrt*.h3, had a significantly lower average level of CQ IC_{50} than both *pfcrt*.h2 (275nmol/L vs 443nmol/L, $p < 6 \times 10^{-8}$) and *pfcrt*.h7 (275nmol/L vs 672nmol/L, $p < 0.005$). The *pfcrt*.h7 haplogroup also had a slightly higher level of chloroquine resistance than *pfcrt*.h2 (672nmol/L vs 443nmol/L, $p < 0.05$), though the number of samples with the *pfcrt*.h7 haplogroup is low. What distinguishes the super-resistant *pfcrt*.h2 and *pfcrt*.h7 haplogroups from the *pfcrt*.h3 haplogroup is the presence of the I356T and N326S mutations (figure 4.9 a). These mutations fall into neighbouring transmembrane domains in the PfCRT protein (figure 4.9 b) and may potentially interact through their physical proximity to produce the perceived increased level of chloroquine resistance.

Differences in the level of chloroquine resistance in field isolates have previously been described for parasites carrying the CVIET and CVIDT haplotypes[90,36], where it was shown that CVIDT parasites exhibit a lower level of CQ IC_{50} than CVIET parasites[90]. The data from the present study show that this observed difference is due to the heterogeneity in CQ IC_{50} levels within the CVIET carry-

ing parasites. Indeed, significant differences in the levels of CQ IC_{50} between CVIDT parasites and archetypical CVIET parasites of haplogroup *pfcrt*.h3 were not seen (259nmol/L vs 275nmol/L, $p > 0.05$) (figure 4.12 a). This is an important observation, because it suggests that the previously observed phenotypic differences between CVIDT and CVIET parasites were not due to differences in their respective haplotypes (*ie.* amino acid positions 72-76), nor due to differences in accessory mutations that accompany these haplotypes (figure 4.3). Rather, the perceived difference was due to the confounding effect of the super-resistant *pfcrt*.h2 and *pfcrt*.h7 haplogroups, which were significantly more chloroquine resistant than CVIDT parasites (443nmol/L & 672nmol/L vs 259nmol/L, $p < 7 \times 10^{-12}$ & $p < 0.005$ respectively).

The *pfcrt*.h2 and *pfcrt*.h7 haplogroups do not only exhibit high-level resistance to chloroquine; the average parasite clearance half-life ($PC_{t1/2}$) for artemisinin is also significantly higher for both of these haplogroups compared to the other *pfcrt* haplogroups ($PC_{t1/2}$ of 4.5hr & 5.1hr vs 2.9hr, $p < 5 \times 10^{-16}$ & $p < 3 \times 10^{-5}$ respectively) (figure 4.12 a). This is not surprising due to the fact that the majority of the *kelch 13* mutations, implicated in artemisinin resistance [13,334,221], have occurred on these two haplogroup backgrounds (figure 4.12 b). Indeed, one of the *pfcrt* mutations defining these haplogroups (I356T) came up as a strong candidate in a genome-wide association study for artemisinin resistance [221]. Furthermore, it is known that mefloquine resistance and piperaquine resistance, mediated by copy number amplifications in the *mdr1* [370,280,282] and *plasmepsin 2/3* [8,372] genes, have occurred largely on *kelch 13* mutant backgrounds [268,9]. Markers for mefloquine and piperaquine resistance were indeed enriched in the samples carrying the super-resistant *pfcrt* haplogroups (figure 4.12 b). However, I did not observe a significantly higher level of mefloquine or piperaquine resistance in these haplogroups (all t-tests: $p > 0.05$) (figure 4.13), likely as a result of the observation that the two

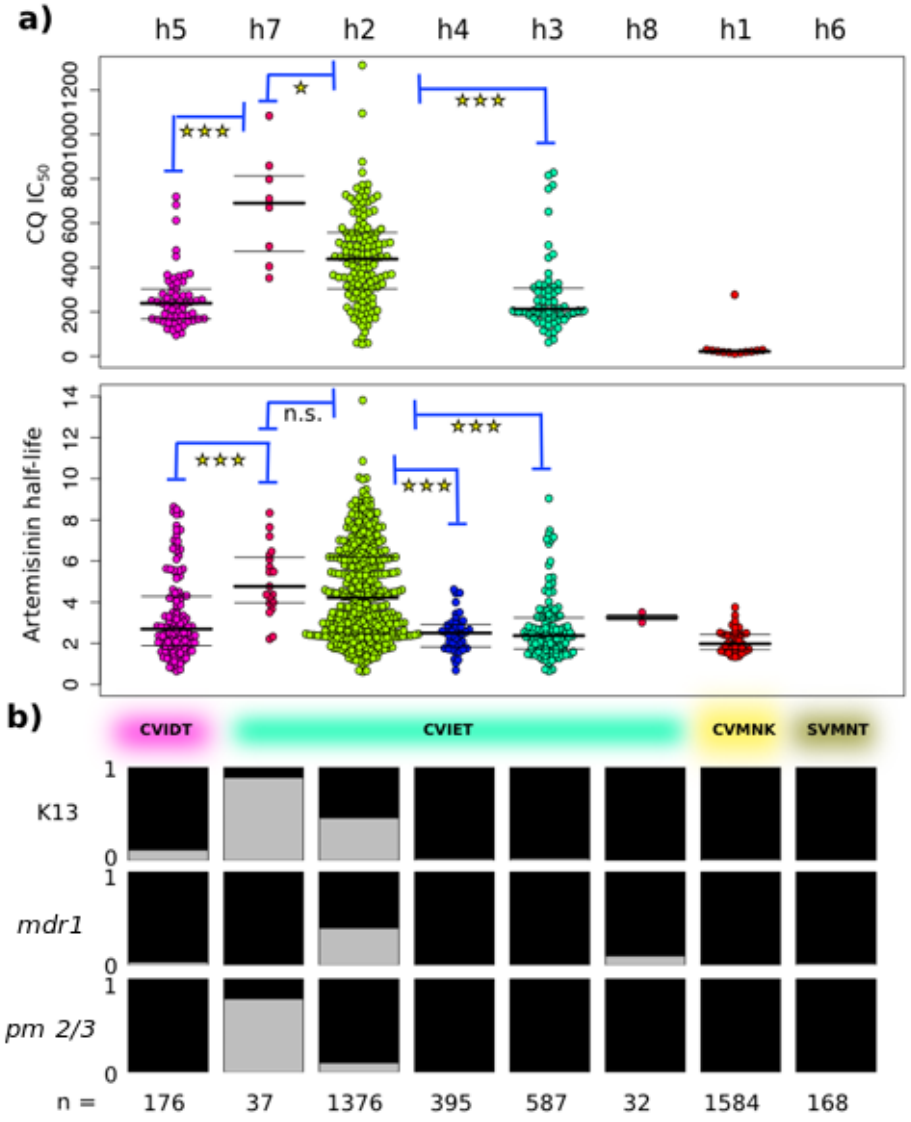


Figure 4.12: Phenotype associations with major pfcrt haplogroups a) Chloroquine IC_{50} by pfcrt haplogroup (above) and artemisinin $PC_{t1/2}$ by pfcrt haplogroup (below). The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (n.s. = $p > 0.05$, * = $p < 0.05$, *** = $p < 0.0005$). b) Additional genetic features of the pfcrt haplogroups, first row showing pfcrt haplotype (positions: 72-76), second row showing proportion of samples with a kelch 13 mutation (grey) and those without (black), third row showing proportion of samples with an mdr1 amplification (grey) and those without (black), and fourth row showing proportion of samples with a plasmepsin 2/3 amplification (grey) and those without (black).

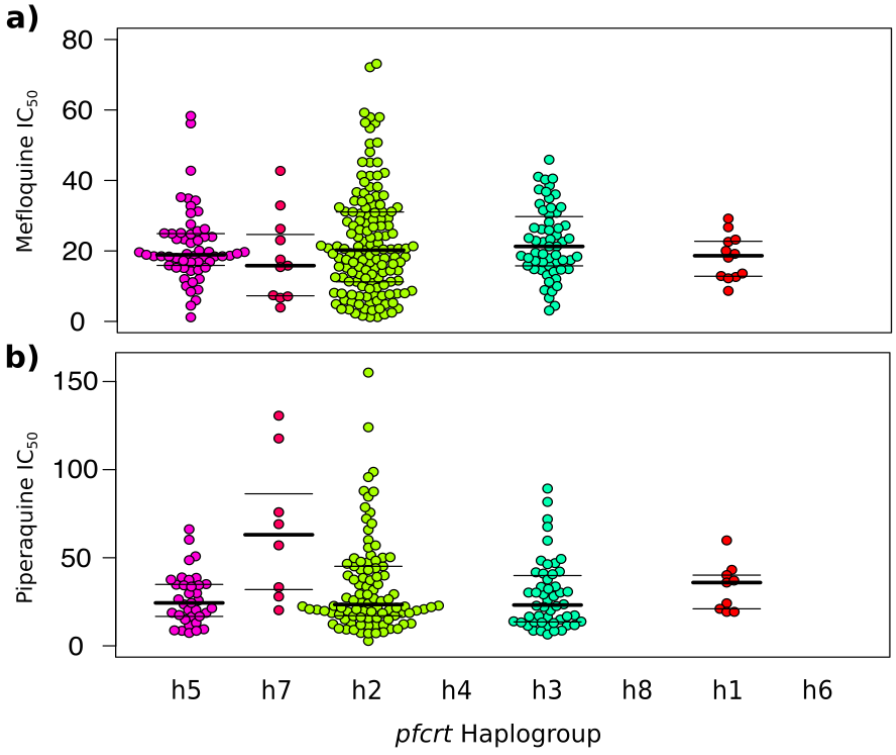


Figure 4.13: a) Mefloquine IC_{50} by pfcrt haplogroup and b) piperaquine IC_{50} by pfcrt haplogroup. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions.

types of copy number amplifications act in an antagonistic manner[8,9]. I have therefore described the presence of two related *pfcrt* haplogroups that exhibit high-level resistance to chloroquine and that act as a genetic backbone for multidrug resistance.

4.8 Distribution and Prevalence of the Super-Resistant Haplogroups

Both of the super-resistant haplogroups, *pfcrt*.h2 and *pfcrt*.h7, are unique to Southeast Asia and are at a very high prevalence in that region (figure 4.10). It is intriguing that, despite its increased chloroquine resistance, *pfcrt*.h2, which is the most common haplogroup in Southeast Asia, isn't known to be found elsewhere in the world. Furthermore, *pfcrt*.h2 has been essentially fixed in western Southeast Asia since at least 2001 (figure 4.14). This suggests that, either, these haplogroups have reduced fitness but ongoing chloroquine pressure in Southeast Asia that sustains them, or that selection for the other drug resistance markers (*ie. kelch 13*, *plasmepsin 2/3*, *mdr1*) maintains these *pfcrt* haplogroups

through hitch-hiking or through some epistatic effects.

Within Southeast Asia, I observed a gradient in the prevalence of the two super-resistant haplogroups (figure 4.14). They are most prevalent in western Southeast Asia, such as Myanmar (96%, 192/200) and Thailand (98%, 722/735), and become less common as you head east within the region across Cambodia (58%, 405/695) to Laos (12%, 12/101) and Vietnam (38%, 66/173). This gradient is particularly evident within Cambodia, where the two haplogroups account for 88% of samples in the western region of Pailin (n = 120), 38% in the central region of Preah Vihear (n = 105), and only 7% in the eastern region of Ratanakiri (n = 144) (figure 4.14). Interestingly, while the pattern within Cambodia is the same, the overall regional pattern is the opposite to that observed for *kelch 13* mutations, which are most common in eastern Southeast Asia and less so in the western parts of the region[219]. The super-resistant *pfcrt* haplogroups therefore likely spread independently of the *kelch 13* mutations, and suggests that they may have originated from the western part of Southeast Asia.

I also observed interesting temporal patterns within Southeast Asia for the prevalence of the *pfcrt*.h2 and *pfcrt*.h7 haplogroups (figure 4.14). In the western parts of the region, their prevalence has remained close to fixation since 2001. However, in Laos and Vietnam I observed an increase in their prevalence from 3% (1/32) to 25% (6/25) (2010-2012) and from 25% (2/8) to 50% (11/22) (2009-2012) across the years of collection, while they appear to have become less common in Cambodia, 67% (4/6) to 38% (18/48) (2008-2014), though there appears to have been a recent spike to 61% (11/18) in 2015 (figure 4.14). The apparent decline in Cambodia is unexpected, especially with the knowledge that the other drug resistance markers have increased in prevalence in Cambodia across the same years of collection[8]. One possible explanation may be the large heterogeneity in prevalence between the different regions in Cambodia, which could cause spurious patterns through uneven sampling across the

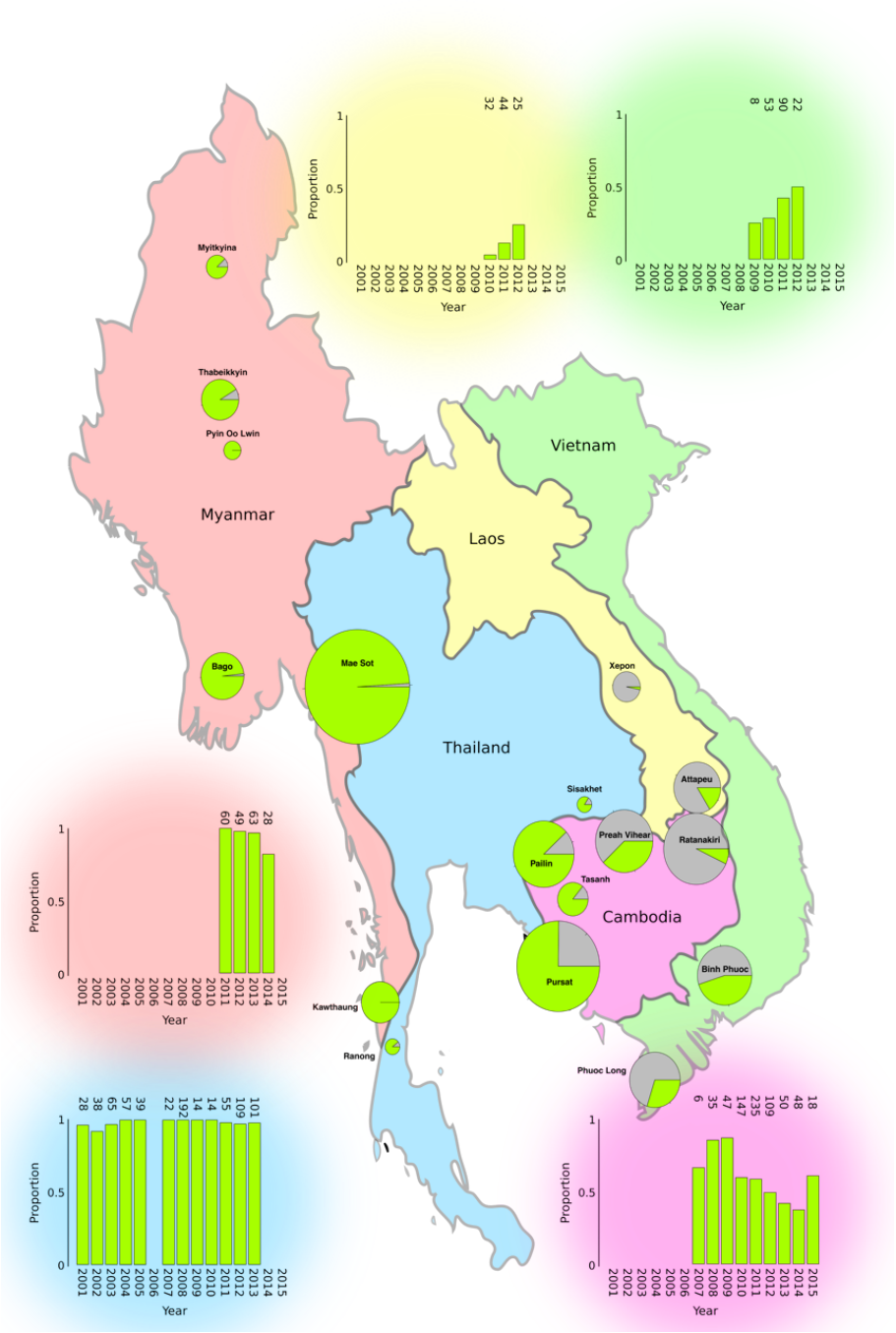


Figure 4.14: Southeast Asian map with pie charts for the different sampling regions showing the proportion of samples that have a super resistant pfcrt haplogroup (pfcrt.h2 and pfcrt.h7), shown in green, compared to other haplogroups, shown in grey. The size of the pie chart is proportional to the number of samples from that region. The bar charts show the proportion of samples that have a super resistant haplogroup by year for the different countries (background colour corresponds to country). The numbers above the bar charts indicate the number of samples collected during that particular year.

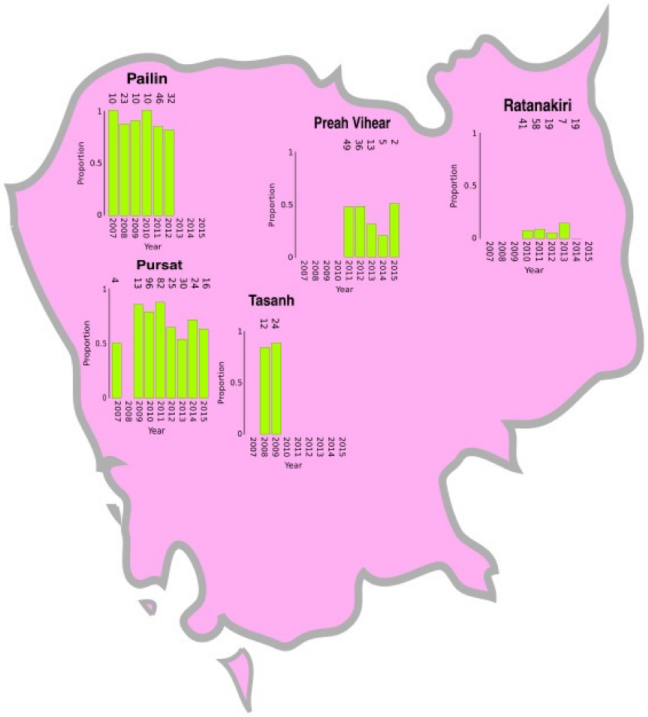


Figure 4.15: Map of Cambodia with bar charts showing the proportion of samples that have a super chloroquine resistant haplogroup (pfcrt.h2 and pfcrt.h7) by year for the different sampling regions. The numbers above the bar charts indicate the number of samples collected during that particular year.

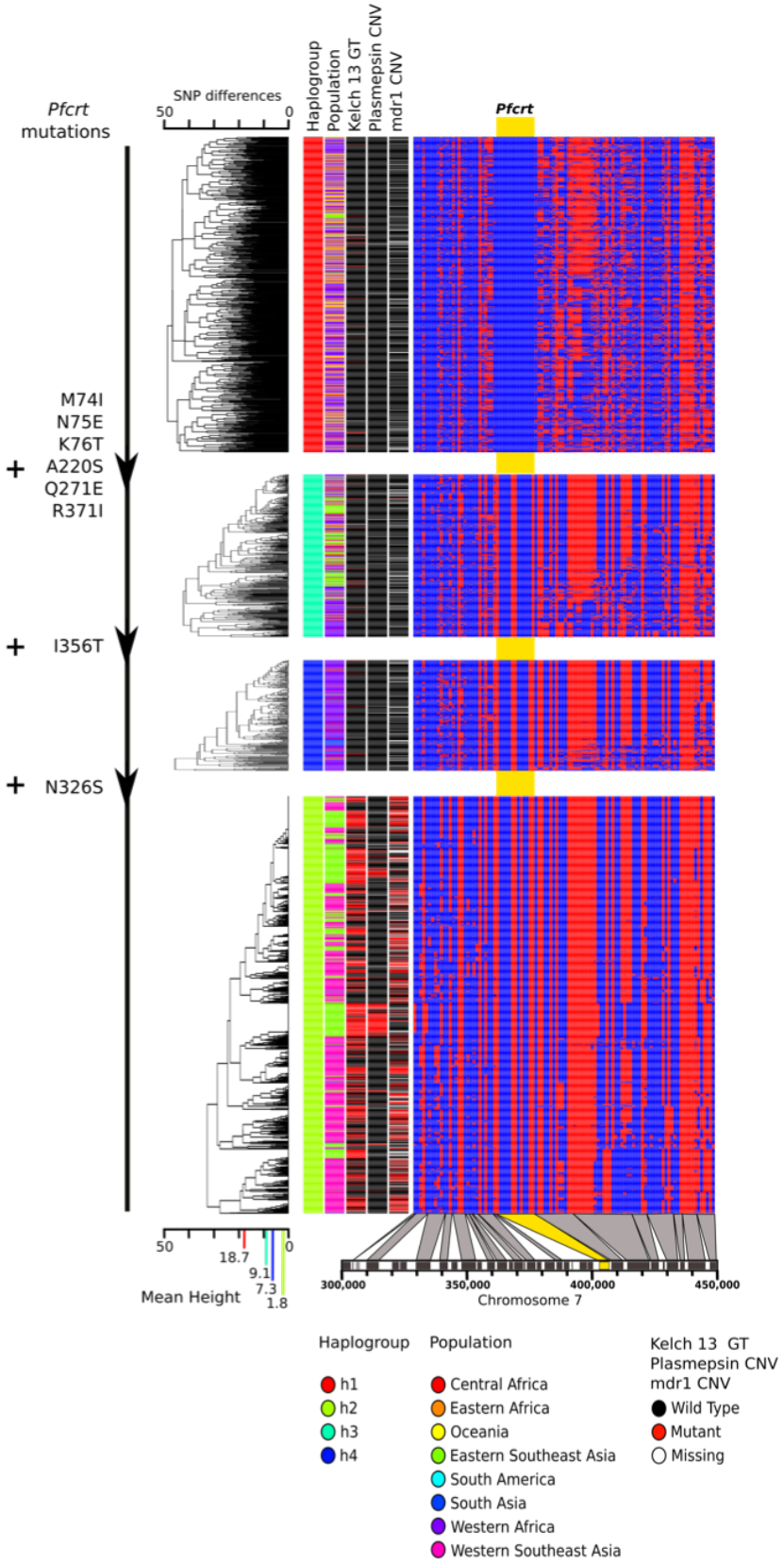
years. Stratifying Cambodia by sampling region, I see that there has not been any change in the prevalence of the super-resistant haplogroups, and that the apparent decrease was due to increased sampling in Pailin and Tasanh in the early years of collection (figure 4.15). On the other hand, the increase in prevalence in Laos and Vietnam is worrying, not only due to the threat of increased chloroquine resistance, but also due to the fact that we know that these haplogroups act as genetic backbones for the other types of drug resistance. The spread of these haplogroups is therefore synonymous with a spread in multidrug resistance and could be an urgent public health issue.

4.9 Strong Linkage Disequilibrium around *pfcrt*

To determine whether these super-resistant haplogroups arose multiple times independently or whether they only have one or a handful of origins, I looked at the coding regions flanking the *pfcrt* gene (102 SNPs, 100kb upstream and 50kb downstream) (figure 4.16). I found that samples with the *pfcrt*.h1

haplogroup, which is identical to the 3D7 wild type reference strain, have comparatively little structure in their flanking regions, with an average of 18.7 SNPs differing between the samples (n = 765). On the other hand, the super-resistant *pfcrt*.h2 haplogroup-containing samples only differ by 1.8 SNPs on average across the same region (n = 1,040), and a striking amount of structure is clearly visible (figure 4.16). The level of structure that is apparent strongly suggests that the *pfcrt*.h2 haplogroup likely only had a small number of origins, possibly only one.

Figure 4.16: Flanking regions around pfcrt by haplogroup, showing variant calls (blue: reference, red: alternate) in coding regions 50kb upstream and 100kb downstream of pfcrt (yellow). Each row corresponding to one sample and each column corresponding to a variant position. The chromosome region is shown below, with variants mapped to the respective genes (black) they are found in. The flanking regions are shown by four pfcrt haplogroups, with the gradual increase of mutations from one haplogroup to the other being shown on the left. Hierarchical clustering trees for each haplogroup show the genetic relationships of the samples to each other based on the flanking regions, with the mean branch height (ie. average number of differences) for each haplogroup shown below. Between the hierarchical trees and the flanking region plot, five columns indicate additional sample-specific information, including: haplogroup membership, P. falciparum population, kelch 13 mutation genotype, plasmepsin 2/3 copy number amplification presence, and mdr1 copy number amplification presence (see legend below for each of these).


Pfcrt
mutations
SNP differences
50
0
Haplogroup
Population
Kelch 13 GT
Plasmepsin CNV
mdr1 CNV
Pfcrt
M74I
N75E
K76T
+ A220S
Q271E
R371I
+ I356T
+ N326S
50
18.7
9.1
7.3
1.8
0
Mean Height
300,000
350,000
400,000
450,000
Chromosome 7
Haplogroup
h1
h2
h3
h4
Population
Central Africa
Eastern Africa
Oceania
Eastern Southeast Asia
South America
South Asia
Western Africa
Western Southeast Asia
Kelch 13 GT
Plasmepsin CNV
mdr1 CNV
Wild Type
Mutant
Missing

An interesting picture emerges when looking at *pfcrt*.h3 and *pfcrt*.h4 which represent potential evolutionary stepping stones from *pfcrt*.h1 to *pfcrt*.h2, as they have all the *pfcrt*.h2 mutations except *pfcrt*.h3 lacking N326S and I356T and *pfcrt*.h4 lacking only N326S. Both *pfcrt*.h3 and *pfcrt*.h4 contain an intermediate amount of SNP differences between the different samples, 9.1 SNPs on average for *pfcrt*.h3 (n = 400) and 7.1 SNPs on average for *pfcrt*.h4 (n = 266), coinciding with their placements between *pfcrt*.h1 and *pfcrt*.h2 (figure 4.16). Surprisingly, the structure that is visible for *pfcrt*.h3 and *pfcrt*.h4 resembles the SNP pattern seen for the *pfcrt*.h2 haplogroup, potentially suggesting that that the three haplogroups have non-independent origins.

As an additional method of investigating the flanking regions, I calculated the extended haplotype homozygosity[303] on either side of the *pfcrt* gene (figure 4.17). This further shows the extent of diversity within the *pfcrt*.h1 samples, as the EHH breaks down extremely fast for this cluster, especially downstream of the gene, where EHH reaches zero almost instantaneously (figure 4.17). This rapid breakdown of EHH potentially suggests the presence of a recombination hotspot downstream of *pfcrt*, which in turn draws attention to the absence of a rapid EHH dropoff in the other three haplogroups. The EHH breakdown for these three haplogroups corresponds to the results obtained by looking at the overall similarity in the flanking regions, *pfcrt*.h3 drops off fastest and *pfcrt*.h2 the slowest (figure 4.17). Calculating the average relative EHH (rEHH)[303] across the 100kb upstream and downstream of *pfcrt* for each of the four tested *pfcrt* haplogroups compared to the other three, I find that *pfcrt*.h1 has the lowest rEHH of 0.04 ± 0.02 (two standard errors of the mean), then *pfcrt*.h3 of 0.82 ± 0.05, *pfcrt*.h4 has a mean rEHH of 2.42 ± 0.12 and *pfcrt*.h2 has the highest rEHH of 15.47 ± 1.23. This further demonstrates the high amount of haplotype conservation around the *pfcrt* locus in the super-resistant haplogroup.

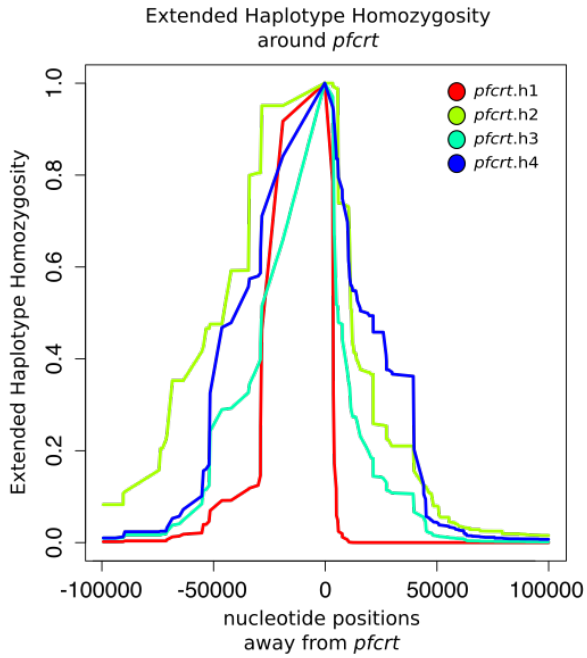


Figure 4.17: Extended haplotype homozygosity (EHH) for 100kb on either side of pfcrt (set at zero) shown for four pfcrt haplogroups (see inset legend). The EHH is the probability that two random samples within a haplogroup are homozygous at all SNPs in the interval from the core region (pfcrt) to the specified distance away from it [303].

4.10 Discussion

The identification of a pair of *pfcrt* haplogroups (*pfcrt*.h2 and *pfcrt*.h7) that act as genetic backbones for multidrug resistance fit within the overall narrative of population substructuring that is emerging within the Southeast Asian setting. It is well-documented that the *P. falciparum* population in Southeast Asia is highly structured [221,8,9], likely a result of the many selective sweeps that have driven through the population as a result of drug pressure. This genetic architecture poses difficulties for traditional genome-wide association studies but also offers a unique opportunity to better understand the evolutionary history of drug resistance. The I356T mutation in *pfcrt*, a marker of *pfcrt*.h2 and *pfcrt*.h7, came up as a confounding candidate SNP in a genome-wide association study of artemisinin resistance [221], which we now know is because it forms part of the *pfcrt* haplogroups that act as genetic backbones for the *kelch 13* mutations. The fact that these haplogroups exhibit super-resistance to chloroquine and that the locus around the *pfcrt* gene is incredibly conserved raise a number of

questions.

The most immediate question is why the mutations involved in artemisinin resistance and other subsequent partner drug resistances occurred on these particular *pfcrt* haplogroups. While the super-resistant haplogroups are essentially fixed in Western Southeast Asia, they are less common in Eastern Southeast Asia where the other drug resistance mutations are thought to have originated from [219,9]. Additionally, we know that *kelch 13* mutations have arisen multiple times independently in the region [342], but seemingly almost exclusively on the super-resistant *pfcrt* haplogroups (figure 4.12 b) (figure 4.16). This situation is further complicated by the observation that parasites with *pfcrt* mutations leading to chloroquine resistance tend to be less fit than wild-type parasites [184]. Therefore, why would we observe *kelch 13* mutations arise multiple times independently only on a specific *pfcrt* haplogroup that is of intermediate prevalence and that likely has reduced fitness compared to other parasites in the region?

One possibility that may explain this seemingly unexpected observation is that there may be some epistatic effects of *pfcrt* on the acquisition of *kelch 13* mutations. This means that the super-resistant *pfcrt* haplogroup is a pre-requisite for the *kelch 13* mutations to be maintained and spread. It is known that mutations in *pfcrt* can affect the expression of a large number of other genes in the genome [326] and it has also been reported that the level of artemisinin resistance conferred by *kelch 13* mutations is dependent on the genetic background of the parasite [334]. The fact that artemisinin is thought to act by inducing oxidative stress that is dependent on hemoglobin digestion by the parasite [168] and the observation that the mutations in *pfcrt* leading to chloroquine resistance result in reduced fitness due to interfering with hemoglobin digestion [190], potentially through reduced heme transport [22], suggest that there may be a mechanistic basis for *pfcrt* mutations providing the optimal background for

artemisinin resistance mutations to continually emerge on. This in turn may also provide an explanation for the lack of artemisinin resistance in Africa, as the super-resistant haplogroups are absent in that region and therefore the *kelch 13* mutations that would arise there would not have the same fitness advantage as those that appear in the Southeast Asian context. Further studies are required to test this hypothesis.

Another question that arises from the results in this study relates to the strong conservation of haplotype structure around the *pfcrt* gene and how this is maintained in the population. The striking similarities in the structure of the flanking regions between the major CVIET containing *pfcrt* haplogroups (*pfcrt*.h2, *pfcrt*.h3, *pfcrt*.h4) suggests that they have had a non-independent origin and that *pfcrt*.h2, which has the highest rEHH, may potentially have arisen from a *pfcrt*.h4 background. Taking into account that *pfcrt*.h3 and *pfcrt*.h4 also exhibit high amounts of structure conservation (figure 4.16) and assuming that they are the result of the spread in chloroquine resistance from Southeast Asia to Africa in the 1970's[12], then it suggests that *pfcrt*.h2 may have a relatively old origin. This is supported by the fact that many of the samples collected in 2001 in Mae Sot contain the *pfcrt*.h2 haplogroup (figure 4.14). One would expect the structure to have broken down across such a long period of time. This is especially true considering the rapid breakdown in EHH I observe in the *pfcrt*.h1 haplogroup containing samples (figure 4.17). It is possible that SNPs in other genes in the vicinity of *pfcrt* may act as compensatory mutations for the *pfcrt* mutations and are therefore maintained, resulting in the observed level of structure. The suggestion that this entire region of the genome is essentially prevented from evolving in order to maintain the *pfcrt* haplogroup implies that there is very strong selection for the *pfcrt* haplogroup.

The origin for this level of selection could be twofold. Firstly, continuing use of chloroquine may se-

lect for high-level chloroquine resistance conferred by the super-resistant haplogroups. Chloroquine is currently still used in the region to treat *Plasmodium vivax* induced malaria cases, potentially creating selective pressure on the *P. falciparum* population in cases of mixed-infections and when people treat themselves using chloroquine obtained illicitly. The observed gradient of *pfcrt*.h2 being most prevalent in Myanmar and the western parts of Thailand (figure 4.14), suggest that the level of selection is stronger in that part of the region and/or that the haplogroup originated there. The implementation of malaria policy is difficult in Myanmar due to widespread poverty and political uncertainty, with counterfeit malaria drugs not being uncommon [129]. It is possible that this may be the source of the super-resistant haplogroups. In addition to this, the appearance of drug resistance mutations to other antimalarial drugs on these super-resistant *pfcrt* backgrounds would have provided a second source of strong selective pressure to maintain these *pfcrt* haplogroups. This is especially true if we assume that there is an actual mechanistic way in which the *pfcrt* haplogroups enable the acquisition of artemisinin resistance as I outlined above.

4.11 Conclusion

Having employed a haplogroup analysis of two well-characterized drug resistance genes, *pfcrt* and *mdr1*, I have identified a pair of *pfcrt* haplogroups that exhibit super-resistance to chloroquine. These *pfcrt* haplogroups, characterized by the combination of a CVIET haplotype with the N326S and I356T mutations, act as genetic backbones for artemisinin resistance and as a consequence also for copy number variations leading to mefloquine and piperaquine resistance. The observation that the *kelch 13* mutations seemingly only persist on these super-resistant *pfcrt* backgrounds suggests a mechanistic inter-relationship, with *pfcrt* acting as a form of prerequisite for artemisinin resistance. These

super-resistant haplogroups are present at high frequency throughout Southeast Asia and may have had a single origin in the western parts of the region as evidenced by the flanking regions surrounding the *pfcrt* gene and the distribution of the haplogroups. The high level of conservation in these flanking regions suggests ongoing selection for these super-resistant haplogroups, potentially through continuing use of chloroquine or as a consequence of selection for resistance to the other antimalarial drugs. The apparent rise in frequency of these haplogroups in Laos and Vietnam is reason for increased vigilance, as it is now clear that these haplogroups act essentially as proxies for multidrug resistance in *P. falciparum*.

I believe it's not only possible to eradicate malaria; I believe it's necessary. Ultimately, the cost of controlling it endlessly is not sustainable. The only way to stop this disease is to end it forever.

Bill Gates, www.gatesnotes.com, 2014 CE

5

Conclusion

5.1 Summary of Results

I set out writing this thesis and performing the associated research with the underlying question of 'What can genomics tell us about human malaria parasites?'. The simple, yet resounding, answer to that question is 'a lot'. Throughout this thesis, I have tackled our current understanding of human malaria and attempted to expand on it from a number of different perspectives. These analyses varied both in the species studied, evolutionary timescale examined, and type and wealth of data that went

into informing the different studies.

In Chapter 1, I took advantage of recent progress in sequencing technology and sample preparation to assemble reference genome sequences for *P. malariae* and for both species of *P. ovale*, thereby filling a crucial gap in our genomic understanding of human malaria. I was able to use those genome sequences to conclusively show that *P. ovale wallikeri* and *P. ovale curtisi* are indeed highly differentiated species. The genome sequences also enabled me to infer the phylogenetic relationship of the different *Plasmodium* species to each other using the largest and most complete amino acid alignment to date, showing that the rodent malaria parasites appear to form a sister clade with the *P. ovale* species and therefore fall within the clade of human-infective malaria parasites, suggestive of an ancestral host switch. I also discovered two large novel gene families in *P. malariae*, *fam-l* and *fam-m*, that occur in doublets throughout the subtelomeres of that species and that appear to be transported to the red blood cell surface where they may adopt an RH5-like fold. Having also assembled the genome sequence of a chimpanzee-infective species closely related to *P. malariae*, *P. malariae*-like, I was able to show symmetry in signals of selection of human-infective species (*P. malariae* and *P. falciparum*) diverging from chimpanzee-infective species (*P. malariae*-like and *P. reichenowi*).

In Chapter 2, harnessing the newly assembled reference genome sequence of *P. malariae*, I characterized a clinical recrudescence case of *P. malariae*. I was able to show that the initial infection (which was the one used for the reference genome assembly), consisted of three haplotypes at different frequencies and that it was the least prevalent haplotype that resulted in the recrudescent infection. Additionally, I discovered that the three haplotypes in the initial infection were very closely related to each other and that they likely resulted from sexual recombination of four parental haplotypes. Comparing SNPs, I identified a number of SNPs in drug resistance genes that may explain how the rare haplotype survived

the drug treatment, while the more prevalent ones were completely cleared. Different scenarios were considered that could explain the recrudescence, with drug resistance providing a potential explanation.

In Chapter 3, I wished to better understand the genetic basis of antimalarial drug resistance and thus focused on the current outbreak of multidrug resistance in Southeast Asia by looking at the response of KEL1/PLA1 *P. falciparum* parasites to mefloquine treatment. I have shown that KEL/PLA1 parasites appear to be hypersensitive to mefloquine treatment compared to wild type parasites due to their acquisition of the PLA1 co-lineage. I have suggested that, as the PLA1 lineage is characterized by the presence of the *plasmepsin 2/3* copy number amplification, that this amplification mediates mefloquine response in an antagonistic way to *mdr1* amplification, mirroring their antagonism on piperaquine response, and suggesting that triple mutants may not necessarily display triple resistance. Additionally, I identified a novel mutation in the *mdr1* gene that appears to have arisen on a *plasmepsin 2/3* amplified background, leading to an even higher level of mefloquine sensitivity. This particular SNP is shown to have had a single origin and has rapidly increased in frequency and spread through the region, all hallmarks of positive selection. Finally, I discussed the potential mechanistic and evolutionary implications of this finding together with the apparent antagonism of the types of copy number amplifications.

In Chapter 4, I used a haplogroup approach to analyse how haplotypes of two well-studied genes, *pfcrt* and *mdr1*, associate with multidrug resistance, doing so by harnessing phenotype data on multiple drugs. I showed that, while I find few associations of drug response with individual *mdr1* haplogroups, except for the one haplogroup containing the novel SNP identified in Chapter 3, that a specific pair of *pfcrt* haplogroups appear to exhibit super-resistance to chloroquine, over and above

the level of resistance conferred by the quintessential chloroquine resistance mutation K76T. These two haplogroups share two important mutations at amino acid positions 326 and 356 that may be functionally mediating this super-resistance. I showed that these super-resistant *pfcrt* haplogroups appear to act as the genetic backbone on which artemisinin and consequently multidrug resistance emerged. These super-resistant haplogroups are restricted to Southeast Asia and display high levels of extended haplotype homozygosity around the *pfcrt* gene. I proposed that the *pfcrt* gene may play a functional role in enabling the acquisition of artemisinin resistance and that the high frequency of these super-resistant *pfcrt* haplogroups in Southeast Asia may be the reason for artemisinin resistance having arisen in that region and not having spread elsewhere.

5.2 An Overarching Narrative

While the overarching question of the thesis was in certain ways a rhetorical one, I attempted to structure and sequence the different analyses in such a way as to add value by exploring the interactions and transferable knowledge that one study can have on another. The most direct example of that is the inability of performing the analysis of clinical recrudescence of *P. malariae* in Chapter 2 without previously having assembled the reference genome sequence for that species in Chapter 1. The analysis of Chapter 1, which in many ways was one of past evolutionary events, now opens up the possibility of improving our understanding of current evolutionary events, such as surveying changes in the neglected human malaria parasite species populations or potentially identifying the emergence of drug resistance early on, as I may have done in Chapter 2. The findings of Chapter 2 are evidently important at the present time in the clinic, but they also bring up questions about more distant evolutionary events, such as whether drug resistance is already present in the field, but simply not recognised, and

whether past clinical recrudescence cases could be explained in a similar way. The identification of KEL1/PLA1 as being hypersensitive to mefloquine sheds light on the evolutionary events that have given rise to the co-lineage, such as the disappearance of *mdr1* amplifications from the field. Finally, while the presence of a super-chloroquine resistant haplogroup acting as a backbone to artemisinin resistance may inform us about the potential sequence of evolutionary events that led to the emergence of artemisinin resistance, it is also an important finding for the present, as it suggests that parasites are very unlikely to revert to becoming chloroquine sensitive in the region.

Another leitfaden throughout this thesis has been that of layering increasingly complex sets of data: I began by comparing genome sequences to each other in Chapter 1, I then layered on clinical metadata in Chapter 2, while in Chapter 3 I added in phenotype data for one drug, and finally, in Chapter 4, I harnessed phenotype data for multiple drugs. At each level, I identified new biology that both expands our understanding of human malaria in general and sheds light on the specifics of antimalarial drug resistance. In Chapter 1, the assembly of the genome sequences led to immediate new findings, such as the identification of new gene families or pseudogenization of RBPs, though many of the results of that chapter were also the consequence of comparing the new assemblies to existing ones of other species, such as improving our understanding of the *Plasmodium* phylogeny, showing the value of additional data in the form of more genome sequences. In Chapter 2, it was only because of the clinical metadata that it became apparent that it was the same patient from whom the two *P. malariae* samples originated from, allowing me to identify the recrudescence of the minor haplotype. The clinical metadata was also crucial in ruling out specific explanations for the recrudescence, giving more weight to the potential explanation of drug resistance. While it was known that there were changes in the frequencies of the *mdr1* and *plasmepsin 2/3* amplifications in the field and that mefloquine appeared to

be effective, it was only with access to the mefloquine IC_{50} values that it was possible to show that the two amplifications appear to be antagonistic and that KEL1/PLA1 parasites are hypersensitive to mefloquine. Finally it was known that certain *pfcrt* mutations were associated with artemisinin resistance, but it was only with phenotype data on multiple drugs that it became apparent that artemisinin and chloroquine resistance were highly correlated in the field because the former appears to have only arisen on a particular *pfcrt* haplogroup mediating chloroquine super-resistance. It has been exciting analysing these datasets and seeing what new questions can be answered with the addition of new types of data.

5.3 Future Directions

There still remains a lot to be learned and the best studies simply raise new questions: how have the two *P. ovale* species speciated? What genes enable *P. malariae* to remain in the host for decades unnoticed? What is the mechanistic explanation for mefloquine hypersensitivity induced by *plasmepsin 2/3* acquisition? Why has multidrug resistance in Southeast Asia exclusively arisen on a single *pfcrt* haplotype? The questions that arise are broad and involve numerous different fields and branches of malaria research; describing or even just listing all the immediate questions is beyond the scope of this conclusion chapter. To focus the discussion, and to keep in line with the two themes guiding the narrative of this thesis, namely the overlaying of increasing levels of data and the interplay between current and past evolutionary events, I wish to outline how I see the next research steps pertaining to each of these.

Data is becoming ever more abundant and complex, as evidenced throughout this thesis. Even a single whole genome sequence contains a mind-baffling amount of information, nevermind analysing

hundreds of them together. In this thesis I have shown how integrating different sets of data can reveal new biological insights. One particular 'combination' of data that I foresee as being incredibly exciting, and that may happen in the coming years, is a pan-*Plasmodium* population genetics study. The jump in my thesis from *P. malariae* and *P. ovale* in Chapters 1 and 2 to *P. falciparum* in Chapters 3 and 4 may appear abrupt and almost *non sequitur*, and in many ways it is because so far these fields of study have not been linked up. As more and more sequencing data becomes available for all human malaria parasite species, as is now happening for *P. vivax* for instance, it will become natural to combine these divergent datasets. Combining these will enable novel analyses to be developed to look for instance at symmetries in selection signals across multiple species, to analyse correlations in changes in transmission intensity (which can be inferred genetically) of different species, or to study the frequency of co-infections and potential genetic adaptations linked to those. At the outset of this thesis, I explained the need for genome sequences for all human malaria parasites in order to enable effective surveillance and a better understanding of those neglected species that will likely pose the most difficulty in the endgame of malaria elimination. Analyses that incorporate timely and well-sampled genetic data from across all human malaria parasites will open up new research avenues that will provide a more holistic picture of malaria elimination, and will ensure that no blind spots remain as we make a concerted push towards the endgoal.

The second exciting field of research that opens up as a result of the increasing availability of data is that of mathematical models based on genomic data. While this thesis explored both past and current evolutionary events and showed how they intersect and influence each other, the application of mathematical models has the potential to inform us about future evolutionary events. Using the knowledge of past and current events in the form of genomic data will enable us to create mathematical mod-

els that utilize that data to make predictions about future evolutionary events. These models do not only rely on high quality genetic data, but also on the data being sampled in a systematic way both temporally and spatially. The Pf6 MalariaGEN dataset that was analysed as part of Chapters 3 & 4 consisted of exactly such data and with ongoing collections, this resource will only continue to become increasingly valuable and information-rich. It is easy to see how these data could theoretically be used to construct transmission models that take into account local drug policies and that are trained on the actual genetic changes in the parasite populations as drug policies change. This is especially true with our improved understanding of the interactions between different drugs, such as the antagonism of piperaquine and mefloquine or how chloroquine super-resistance appears to be the genetic backbone for artemisinin resistance. Policy-makers want to know what will happen next when they switch the recommended treatment. While in the past, traditional transmission models were used to inform these choices, they were always parameterized with assumptions that were often difficult to support objectively. By supplanting these textbook parameters with proxies that are accurately inferred from genetic information will enable models to be parameterized according to the specific scenario and also allow the parameters to be continually updated as new genetic data comes in from the field.

These future advancements in research will play pivotal roles as we approach near-elimination settings. Bringing together the genomic epidemiological models with the pan-*Plasmodium* perspective I outlined earlier is, in my opinion, the holy grail of genomic surveillance for malaria elimination. The findings that I have presented throughout this thesis have laid in different ways the foundation for these future breakthroughs. While many of the results contained within this thesis have direct applications in the field at present, they also pave the way for future discoveries to be made that will have the greatest impact on our battle against malaria.

A

Chapter 1 Methods

A.1 Co-infection Mining

I aligned the *P. malariae* (AB354570) and *P. ovale* (AB354571) mitochondrial genome sequences against those of *P. falciparum* [115], *P. vivax* [51], and *P. knowlesi* [266] using MUSCLE [94]. For each species, I identified three 15bp stretches within the *cox1* gene that contained two or more species-specific SNPs. I searched for these 15bp species-specific barcodes within the sequencing reads of all 2,512 samples from the Pf3K global collection (www.malariagen.net). Samples that contained at least two sequenc-

ing reads matching one or more of the 15bp barcodes for a specific species were considered to be positive for that species (Chapter 1: table 1.1). I found good correspondence between the three different barcodes for each species, with over 80% of positive samples being positive for all three barcodes. I generated pseudo-barcodes by changing two randomly selected nucleotide bases at a time for 10 randomly selected 15bp region in the *P. vivax*[51] mitochondrial genome. I did not detect any positive hits using these pseudo-barcodes. As an additional negative control, I searched for *P. knowlesi* co-infections, but did not find any samples positive for this species. Two samples (PocGH01, PocGH02) had high numbers for all three *P. ovale* barcodes and were used for reference genome assembly and SNP calling respectively.

A.2 Parasite Material

All *P. ovale* samples were obtained from symptomatic patients diagnosed with a *P. falciparum* infection. The two *P. o. curtisi* samples (PocGH01, PocGH02) identified through co-infection mining (see above), were from two patients testing positive on a CareStart® (HRP2 based) rapid malaria diagnostic test (RDT) kit at the Navrongo War Memorial hospital, Ghana. One *P. o. wallikeri* sample (PowCR01) and one *P. o. curtisi* sample were from uncomplicated malaria patients testing positive by light microscopy at the Mile 16 - Bolifamba Health Centre, Buea, Cameroon. The other *P. o. wallikeri* sample (PowCR02) was obtained from an individual with asymptomatic parasitemia enrolled through a community survey in Mutengene, Cameroon. For all samples, following consent obtainment, about 2-5mls of venous blood was obtained and then diluted with one volume of PBS. This was passed through CF11 cellulose powder columns to remove leucocytes prior to parasite DNA extraction.

The two *P. malariae*-like samples, PmlGA01 and PmlGA02, were extracted from Chimpanzee blood obtained during routine sanitary controls of animals living in a Gabonese sanctuary (Park of La Lékédi, Gabon). Blood collection was performed following international rules for animal health. Within six hours after collection, host white blood cell depletion was performed on fresh blood samples using the CF11 method[21]. After DNA extraction using the Qiagen blood and Tissue Kit and detection of *P. malariae* infections by Cytb PCR and sequencing[253], the samples went through a whole genome amplification step[263].

One *P. malariae* sample, PmMA01, was collected from a patient with uncomplicated malaria in Faladje, Mali. Venous blood (2–5mL) was depleted of leukocytes within 6 hours of collection as previously described[161]. The study protocol was approved by the Ethics Committee of Faculty of Medicine and Odontomatology and Faculty of Pharmacy, Bamako, Mali.

Four samples of *P. malariae* were obtained from travellers returning to Australia with malaria. PmUG01 and PmID01 were sourced from patients returning from Uganda and Papua Indonesia respectively, who presented at the Royal Darwin Hospital, Darwin, with microscopy-positive *P. malariae* infection. PmMY01 was sourced from a patient presenting at the Queen Elizabeth Hospital, Sabah, Malaysia, with microscopy-positive *P. malariae* infection. Patient sample PmGN01 was collected from a patient who presented to Royal Brisbane and Womens Hospital in 2013 on return from Guinea.

Venous blood samples were subject to leukodepletion within 6 hours of collection. PmUG01 was leukodepleted using a commercial Plasmodipur filter (EuroProxima, The Netherlands); home-made cellulose-based filters were used for PmID01 and PmMY01, while PmGN01 was leukodepleted using an inline leukodepletion filter present in the venesection pack (Pall Leukotrap; WBT436CEA). DNA

extraction was undertaken on filtered blood using commercial kits (QIAamp DNA Blood Midi kit, Qiagen Australia).

For samples PmUG01, PmID01 and PmMY01, ethical approval for the sample collection was obtained from the Human Research Ethics Committee of NT Department of Health and Families and Menzies School of Health Research (HREC-2010-1396 and HREC-2010-1431) and the Medical Research Ethics Committee, Ministry of Health Malaysia (NMRR-10-754-6684). For sample PmGN02, ethical approval was obtained from the Royal Brisbane and Womens Hospital Human Research Ethics Committee (HREC/10/QRBW/379) and the Human Research Ethics Committee of the Queensland Institute of Medical Research (p1478).

A.3 Sample Preparation and Sequencing

One *P. malariae* sample, PmUG01, was selected for long read sequencing, using Pacific Biosciences (PacBio), due to its low host contamination and abundant DNA. Passing through a 25mm blunt-ended needle, 6ug of DNA was sheared to 20-25kb. SMRT bell template libraries were generated using the PacBio issued protocol (20kb Template Preparation using the BluePippin™ Size-Selection System). After a greater than 7kb size-selection using the BluePippin™ Size-Selection System (Sage Science, Beverly, MA), the library was sequenced using P6 polymerase and chemistry version 4 (P6/C4) in 20 SMRT cells (table A.1).

The remaining isolates were sequenced with Illumina Standard libraries of 200-300bp fragments and amplification-free libraries of 400-600bp fragments were prepared[43] and sequenced on the Illumina HiSeq 2000 v3 or v4 and the MiSeq v2 according to the manufacturer's standard protocol (table A.1). Raw sequence data was deposited in the European Nucleotide Archive (table A.1).

Table A.1: Origin, sequencing statistics, and usage of samples in this study

Species	Sample ID	Accession ID	Origin	Sequencing Platform	Read Length	Usage	Library type
P. malariae	PmUG01	ERS1110315	Uganda	PacBio RS II P6/C4	N/A	Reference Genome	Size-selected
P. malariae	PmMY01	ERS1110317	Malaysia	Illumina MiSeq v2, Illumina HiSeq 2000 v4	150bp PE 125bp PE	SNP Calling	Amplification free
P. malariae	PmID01	ERS1110321	Papua Indonesia	Illumina MiSeq v2, Illumina HiSeq 2000 v4	150bp PE 125bp PE	SNP Calling	Amplification free
P. malariae	PmMA01	ERS1110325	Mali	Illumina MiSeq v2	150bp PE	SNP Calling	Standard
P. malariae	PmGN01	ERS567899	Guinea	Illumina MiSeq v2	150bp PE 250bp PE	SNP Calling	Standard
P. malariae-like	PmlGA01	ERS434571	Gabon	Illumina MiSeq v2	150bp PE	Draft Genome	WGA* - Amplification free
P. malariae-like	PmlGA02	ERS434565	Gabon	Illumina MiSeq v2	150bp PE	SNP Calling	WGA* - Amplification free
P. ovale curtisi	PocGH01	ERS013096	Ghana	Illumina HiSeq 2000 v3	100bp PE	Reference Genome	Standard
P. ovale curtisi	PocGH02	ERS360497	Ghana	Illumina HiSeq 2000 v3	100bp PE	SNP Calling	Standard
P. ovale curtisi	PocCR01	ERS418861	Cameroon	Illumina HiSeq 2000 v3	100bp PE	SNP Calling	Standard
P. ovale wallikeri	PowCR01	ERS418894	Cameroon	Illumina HiSeq 2000 v3	100bp PE	Draft Genome	Standard
P. ovale wallikeri	PowCR02	ERS418932	Cameroon	Illumina HiSeq 2000 v3	100bp PE	SNP Calling	Standard

*WGA: Whole Genome Amplification

A.4 Genome Assembly

The PacBio sequenced *P. malariae* sample, PmUG01, was assembled using HGAP[59] with an estimated genome size of 100Mb to account for the host contamination (≈85% Human). The resulting assembly was corrected initially using Quiver[59], followed by iCORN[260]. PmUG01 consisted of two haplotypes, with the majority haplotype being used for the iCORN[260], and a coverage analysis was performed to remove duplicate contigs. Additional duplicated contigs were identified using a BLASTN[5] search, with the shorter contigs being removed if they were fully contained within the longer contigs or merged with the longer contig if their contig ends overlapped. Host contamination was removed by manually filtering on GC, coverage, and BLASTN hits to the non-redundant nucleotide database[5].

The Illumina based genome assemblies for *P. o. curtisi*, *P. o. wallikeri*, and *P. malariae*-like were performed using MaSURCA[383] for samples PocGH01, PowCR01, and PmlGA01 respectively. To confirm that the assemblies were indeed *P. ovale*, I mapped existing *P. ovale* capillary reads to the assemblies (www.ncbi.nlm.nih.gov/Traces/trace.cgi?view=search). Prior to applying MaSURCA[383], the samples were mapped to the *P. falciparum* 3D7 reference genome[115] to remove contaminating reads. The draft assemblies were further improved by iterative uses of SSPACE[37], GapFiller[236] and IMAGE[350]. The resulting scaffolds were ordered using ABACAS[18] against the *P. vivax* PVP01[20] assembly (both *P. ovale*) or against the *P. malariae* PacBio assembly (*P. malariae*-like). The assemblies were manually filtered on GC, coverage, and BLASTN hits to the non-redundant nucleotide database[5]. iCORN[260] was used to correct frameshifts. Finally, contigs shorter than 1 kilobase (kb) were removed.

Table A.2: Accession numbers for the assembled genome sequences

Assembly Name	Study ID	Sample ID	Contig Accession	Chromosome Accession
PmUG01	PRJEB2579	ERS1110315	FLRL01000001-FLRL01000047	LT594622-LT594637
PocGH01	PRJEB2579	ERS013096	FLRI01000001-FLRI01000638	LT594582-LT594597
PowCR01	PRJEB2579	ERS418894	FLRJ01000001-FLRJ01000771	LT594505-LT594520
PmlGA01	PRJEB2579	ERS434571	FLRK01000001-FLRK01000035	LT594489-LT594503

Using two more samples, PocGH02 and PowCR02, additional draft assemblies of both *P. ovale* species were produced using MaSURCA[383] followed by RATT[257] to transfer the gene models from the high-quality assemblies.

The genome sequences and annotations are currently available on GeneDB (www.genedb.org) and the genome sequences have been deposited into the European Nucleotide Archive (www.ebi.ac.uk/ena). Accession numbers for all reads generated for this study can be found in table A.1 . Accession IDs for the assembled genome sequences can be found in table A.2.

A.5 Gene Annotation

RATT[257] was used to transfer gene models based on synteny conserved with other sequenced *Plasmodium* species (*P. falciparum*[115], *P. vivax*[51], *P. berghei*[256], and *P. gallinaceum*[38]). In addition, genes were predicted *ab initio* using AUGUSTUS[333], trained on a geneset consisting of manually curated *P. malariae* and *P. ovale* genes respectively. Ulrike Boehme from the Wellcome Sanger Institute identfied non-coding RNAs and tRNAs using Rfam 12.0[238] and then curated gene models for both the *P. malariae* and *P. o. curtisi* reference genomes, using Artemis[300] and the Artemis Comparison Tool (ACT)[55]. She also used these tools to manually identify deleted and disrupted genes.

A.6 Phylogenetics

Following ortholog assignment using BLASTP[5] and OrthoMCL[194], amino acid sequences of 1,000 core genes from 12 *Plasmodium* species (*P. gallinaceum*[38], *P. falciparum*[115], *P. reichenowi*[259], *P. knowlesi*[266], *P. vivax*[20], *P. cynomolgi*[341], *P. chabaudi*[256], *P. berghei*[256], and the four assemblies produced in this study) were aligned using MUSCLE[94]. The alignments were cleaned using GBlocks[343] with default parameters to remove non-informative and gapped sites. The cleaned non-zero length alignments were then concatenated. This resulted in an alignment of 421,988 amino acid sites per species. The optimal substitution model for each gene partition was determined by running RAxML[332] for each gene separately using all implemented substitution models. The substitution models that generated the tree with the highest likelihood were used for each gene partition. A maximum likelihood phylogenetic tree was constructed using RAxML[332] with 100 bootstraps[331] (Chapter 1: figure 1.1 b). To confirm this tree, I utilized different phylogenetic tools including PhyloBayes[182] and PhyML[128], a number of different substitution models within RAxML[332], starting the tree search from the commonly accepted phylogenetic tree, and removing sites in the alignment which supported significantly different trees. All approaches yielded the final tree reported in Chapter 1 with the highest likelihood. Figtree was used to colour the tree (http://tree.bio.ed.ac.uk/software/figtree/).

A phylogenetic tree of four *P. malariae* (PmID01, PmGN01, PmGN02, PmMY01) and all *P. malariae*-like samples (PmlGA01, PmlGA02) was generated using PhyML[128] based on all *P. malariae* genes. For each sample, the raw SNPs as called using the SNP pipeline (see below), were mapped onto all genes to morph them into sample specific gene copies using BCFtools[192]. Amino acids for all genes were concatenated and cleaned using GBlocks[343].

A.7 Divergence Dating

Species divergence times were estimated using the Bayesian inference tool G-PhoCS[125], a software which uses thousands of unlinked neutrally evolving loci and a given phylogeny to estimate demographic parameters. One additional sample per assembly (PmGN01 for *P. malariae*, PocGH02 for *P. o. curtisi*, PowCR02 for *P. o. wallikeri*, and PmlGA02 for *P. malariae*-like) was used to morph the respective assembly using iCORN[260]. Regions in the genomes without mapping were masked, as iCORN[260] would not have morphed them. Unassigned contigs and subtelomeric regions were removed for this analysis due to the difficulty of alignment. Repetitive regions in the chromosomes of the four assemblies and the four morphed samples were masked using DUSTmasker[227] and then the chromosomes were aligned using FSA[42]. The *P. o. wallikeri* and the *P. o. curtisi* chromosomes were aligned against each other, as were the *P. malariae* and *P. malariae*-like chromosomes. The alignments were split into 1kb loci, removing those that contained gaps, masked regions, and coding regions to conform with the neutral loci assumption of G-PhoCS[125]. G-PhoCS[125] was run for one million MCMC-iterations with a sample-skip of 1,000 and a burn-in of 10,000 for each of the two species pairs. Follow-up analyses using Tracer (http://beast.bio.ed.ac.uk/Tracer) confirmed that this was sufficient for convergence of the MCMC chain in all cases. In the model, I assumed a variable mutation rate across loci and allowed for on-going gene flow between the populations. The tau values obtained from this were 0.0049 for *P. malariae* and 0.0434 for *P. ovale*.

The tau values were used to calculate the date of the split, using the formula (tau x G)/mu, where G is the generation time in years and mu is the mutation rate. By assuming a generation time of 65 days[77], I estimated a mutation rate of approximately 3.8 x 10^{-10} SNPs/site/year by optimizing the *P.*

falciparum/*P. reichenowi* split to 4 million years ago, a date that was estimated previously[322]. For *P. malariae*, a generation time of 100 days was used to account for the longer intra-erythrocytic cycle.

A.8 3D Structure Prediction

The I-TASSER[380] version 4.4 online web server[381] (zhanglab.ccmb.med.umich.edu/I-TASSER) was used for 3D protein structure prediction. Predicted structures with a TM-score of over 0.5 were considered reliable as suggested in the I-TASSER user guidelines[377]. TM-align[382], as implemented in I-TASSER[381], was used to overlay the predicted protein structure with existing published protein structures.

A.9 Hypnozoite Gene Search

Using the OrthoMCL[194] clustering between all sequenced *Plasmodium* species used for the phylogenetic analysis (see above), I examined clusters containing only *P. vivax* P01 genes, *P. cynomolgi*[341] genes and genes of both of the *P. ovale* species. Additionally, I examined *P. o. curtisi* orthologs of previously published hypnozoite gene candidates[341], looking in the 1kb 5' upstream region for any of the four ApiAP2 motifs[47] involved in sporozoite regulation and expression: GCATGC (PF3D7_1466400), GCCCCG (PF3D7_1342900), TAAGCC (PF3D7_1342900), and TGTTAC (PF3D7_0420300).

A.10 Gene Family Analysis

All *P. malariae*, *P. ovale*, and *P. vivax* P01 genes were compared in a pairwise manner using BLASTP[5], with genes having a minimum local BLAST hit of 50% identity over 150 amino acids or more be-

ing considered connected. These gene connections were visualized in Gephi[27] using a Fruchterman-Reingold[109] layout and with unconnected nodes removed.

P. malariae, *P. o. curtisi* and *P. o. wallikeri* protein sequences for *Plasmodium* interspersed repeat (*pir*) genes, excluding pseudogenes, were combined with those from *P. vivax* P01 , *P. knowlesi*[266], *P. chabaudi* AS v3 (genedb.org/Homepage/Pchabaudi), *P. yoelii* 17X v2[256], and *P. berghei* v3 (genedb.org/Homepage/Pberghei). Adam Reid from the Wellcome Sanger Institute used these sequences to cluster them using tribeMCL[96] with blast E-value 0.01 and inflation 2, resulting in 152 subfamilies. He then excluded clusters with one member and plotted the number of genes per species in each subfamily in a heatmap using the heatmap.2 function in ggplots in R-3.1.2.

The *pir* genes from two *P. o. curtisi* and two *P. o. wallikeri* assemblies (two high-quality and two draft assemblies) were compared in a pairwise manner using BLASTP[5] with a 99% identity over a minimum of 150 amino acids cutoff. The gene-gene connections were visualized in Gephi[27] using a Fruchterman-Reingold[109] layout after removing unconnected nodes.

A.11 Mirror Tree Analysis

Using Artemis[300], 79 *fam-m* and *fam-l* doublets that were confidently predicted as being paired-up were manually selected based on their dispersal throughout the subtelomeres of different chromosomes. The Mirrortree[248] web server (http://csbg.cnb.csic.es/mtserver/) was used to construct mirror trees for these 79 doublets. Of these, 35 doublets with recent branching from another doublet were manually selected to enrich for genes under recent selection. To control for chance signals of co-evolution based on their subtelomeric location, the same methodology was repeated by choosing 79 *pir* genes in close proximity of *fam-m* genes as 'pseudo-doublets' and paired up in the Mirrortree[248]

web server.

A.12 Reticulocyte Binding Protein (RBP) Phylogenetic Plot

Full-length RBP genes were manually inspected using ACT[55] and verified to either be functional or pseudogenized by looking for sequencing reads in other samples that confirm mutations inducing premature stop codons or frameshifts. All functional RBPs were aligned using MUSCLE[94] and cleaned using GBlocks[343]. PhyML[128] was used to construct a phylogenetic tree of the different RBPs. Figtree was used to colour the tree (http://tree.bio.ed.ac.uk/software/figtree/).

A.13 SNP Calling

Additional *P. malariae* (PmMY01, PmID01, PmMA01, PmGN01) and *P. o. curtisi* (PocGH01, PocGH02, PocCR01) samples were mapped back against the reference genomes using SMALT (-y 0.8, -i 300). As outgroups, *P. malariae*-like (PmlGA01, PmlGA02) and *P. o. wallikeri* (PowCR01, PowCR02) were also mapped against the *P. malariae* and *P. o. curtisi* genomes respectively. The resulting bam format files were merged for either of the two genomes, and GATK's[215] Unified Genotyper was used to call SNPs from the merged bam format files. Per GATK's[215] best practices, SNPs were filtered by quality of depth (QD > 2), depth of coverage (DP > 10), mapping quality (MQ > 20), and strand bias (FS < 60). Additionally, all sites with missing data for any of the samples or with heterozygous calls were filtered away. Finally, I filtered away sites that were masked using DUSTmasker[227] to remove repetitive and difficult to map regions. The same methodology was also applied to two *P. vivax* samples (SRR3400910 & SRR332566) and two *P. falciparum* Pf3K field samples (PF0066-C & PF0038-C) for comparative purposes.

A.14 Molecular Evolution Analysis

To calculate the nucleotide diversity for the different species, I extracted all filtered SNPs in the genomes excluding the subtelomeres. I then counted the number of pairwise differences between the different samples divided by the resulting genome size, comprising three comparisons for species with three samples (*P. malariae*, *P. o. curtisi*, *P. vivax*, *P. falciparum*) and one comparison for species with two samples (*P. o. wallikeri*, *P. malariae*-like). These estimates were then averaged by species.

The filtered SNPs were used to morph the reference genomes using BCFtools[192] for each sample, from which sample-specific gene models were obtained. Nucleotide alignments of each gene were then generated. Codons with alignment positions that were masked using DUSTmasker[227] were excluded. For each alignment (*ie.* gene), I calculated HKA[152], MK[174], and Ka/Ks[241] values (see below). Subtelomeric gene families and pseudogenes were excluded from the analysis. The results were analysed and plotted in RStudio (http://www.rstudio.com/).

The first measure of selection I calculated for each gene was the Hudson-Kreitman-Aguade ratio (HKAr)[152]. HKAr is the ratio of interspecific nucleotide divergence to intraspecific polymorphisms, it is thereby a measure of adaptive evolution, *ie.* an overrepresentation of recent polymorphisms compared to the expected 'neutral' rate implied by the interspecific nucleotide divergence. To calculate the HKAr, I counted the proportion of pairwise nucleotide differences intra-specifically (*ie.* within *P. malariae* and within *P. o. curtisi*) and inter-specifically (*ie.* between *P. malariae* and *P. malariae*-like, between *P. o. curtisi* and *P. o. wallikeri*). The intraspecific comparisons were averaged to get the genes' nucleotide diversity 'π' and these were divided by the average interspecific comparisons, the nucleotide divergence 'K'. The HKAr is therefore π/K for each gene.

The second measure of selection to be calculated was the McDonald Kreitman (MK) skew[174]. The MK skew is a measure of maintained polymorphisms, *ie.* an overrepresentation of polymorphic non-synonymous mutations compared to fixed ones (relative to the ratio polymorphic to fixed synonymous changes). The MK skew was calculated for each gene by obtaining the number of fixed and polymorphic changes, as well as a corresponding p-value using a previously described software[143]. Specifically, the skew was calculated as $\log_2(((N_{poly}+1)/(S_{poly}+1))/((N_{fix}+1)/(S_{fix}+1)))$ where N_{poly} and N_{fix} are polymorphic and fixed non-synonymous substitutions respectively, while S_{poly} and S_{fix} refer to the polymorphic and fixed synonymous substitutions respectively.

The final measure of selection I calculated was the average Ka/Ks ratio[241] for each gene. The Ka/Ks ratio, also known as the d_n/d_s ratio, is the ratio of nonsynonymous to synonymous changes, thereby being a measure of positive (or negative) selection if there is an over- (or under)-representation of nonsynonymous changes. I took the cleaned alignments of the above MK skew, extracting the pairwise sequences of *P. malariae* and *P. malariae*-like (and of *P. o. curtisi* and *P. o. wallikeri*). The Ka/Ks values for each pair were calculated as per the method introduced in[240] as implemented in the Bio::Align::DNAStatistics module, averaging across samples within a species. Briefly, in the[240] method, for a set of *r* codons in a gene, the proportion of synonymous (p_s) and nonsynonymous(p_n) mutations are calculated as S_d/S and N_d/N respectively, where S_d and N_d are the sums of the number of respective differences per codon (represented as the probability of the different mutational pathways that could result in the codon change), while S and N are the average number of synonymous and nonsynonymous sites compared. Finally, from p_s and p_n we can calculate d_s and d_n (which give the d_n/d_s ratio) using the formula: $d = -(3/4) * \log_e(1-(4/3) * p)$.

Using existing RNA-Seq data from seven different life-cycle stages in *P. falciparum*[203], sequencing

reads were mapped against spliced gene sequences (exons, but not UTRs) from the *P. falciparum* 3D7 reference genome[115] using Bowtie2[179] v2.1.0 (-a -X 800 –x). Read counts per transcript were estimated using eXpress v1.3.0[294]. Genes with an effective length cutoff below 10 in any sample were removed. Summing over transcripts generated read counts per gene. Numbers were averaged for all gametocyte stages and for all blood stages. Genes with no stage having 10 or more reads were classified as being expressed elsewhere. Genes in *P. malariae* and *P. ovale* were classified by their *P. falciparum* ortholog's maximum expression stage if the difference between the maximum expression stage and the second highest stage was larger than the difference between the second and third highest stage, otherwise the gene was classified as having no peak expression.

The GO term enrichment analysis was performed in R, using TopGO[3]. As a GO-database, the predicted GO terms from the *P. falciparum* 3D7 genes orthologous to the *P. malariae* and *P. o. curtisi* genes included in the analysis were used.

B

Additional Phylogenetics

B.1 Tree Sensitivity Testing

A number of conflicting phylogenetic trees of the *Plasmodium* genus have been published that differ in their placement of *P. ovale* and *P. malariae*. The two most commonly reported topologies either place *P. ovale* as a sister taxon to the rodent malaria parasites [14] (Tree A) or as an outgroup to *P. malariae* and *P. vivax* [339,308,10] (Tree B). The same studies also place *P. malariae* as either a distant outgroup to both the rodent malaria parasites and the *P. vivax* clade [14], or as being more closely related to *P. vivax*

than *P. ovale*, with *P. malariae* thereby being a close outgroup to the primate infective clade[308,10]. A recent study using draft genome sequences supported Tree B[10], while the phylogenetic tree presented here supports the Tree A topology (Chapter 1: figure 1.1 b).

To better understand the origin of these conflicting reports and to showcase how robust my present analysis is, it is important to delve into the specifics of building the phylogenetic trees. As described in the methods section in Appendix A, following orthologue assignment using BLASTP[5] and OrthoMCL[194], amino acid sequences of 1,000 core genes from 12 *Plasmodium* species (*P. gallinaceum*[38], *P. falciparum*[115], *P. reichenowi*[259], *P. knowlesi*[266], *P. vivax*[20], *P. cynomolgi*[341], *P. chabaudi*[256], *P. berghei*[106], and the four assemblies produced here) were aligned using MUSCLE[94]. The alignments were cleaned using GBlocks[343] with default parameters to remove non-informative and gapped alignment columns. The cleaned non-zero length alignments were then concatenated. This resulted in an alignment of 421,988 amino acid sites per species, the largest such alignment including *P. malariae* and *P. ovale* used to date.

I determined the sensitivity of the tree topology to different parameters and tree-building algorithms, I analysed the 1,000 core gene alignment used for the tree reported in Chapter 1 using a number of different algorithms for phylogenetic inference. I utilized different tree-building softwares, including RAxML[332] (see below), PhyloBayes[182] (using ratecat, cat, and uni models), and PhyML[128] (LG model with optimized site rates), all resulting in Tree A. I also used a number of different amino acid substitution models implemented within RAxML[332] version 8.2.4., including JTT, LG, LG4M, LG4X, GTR_unlinked, GTR, and DAYHOFF. All these models were tested with both CAT and GAMMA site distribution rates. All substitution models resulted in Tree A. In addition to this, I calculated the optimal substitution model for each gene partition by running RAxML[332] for each gene

separately using all implemented substitution models (minimum AIC). Using this substitution model optimized partitioned alignment I still generated Tree A. In order to determine whether the conflicting topology would be a local optimum and was therefore not found using my other approaches, I used Tree B as the starting tree using RAxML[332] with a PROTGAMMAJTT substitution model. This approach still converged on Tree A, indicating that my alignment does not support the Tree B topology.

I performed bootstrapping as implemented within RAxML[331], obtaining very good bootstrap support for all nodes (Chapter 1: figure 1.1 b). I also tested the RAxML '-f S' parameter with a window size of 10 and Tree A as the reference tree. This computes phylogenetic signal strengths for each site in the alignment using a leave-one-out approach[29]. I used this output to filter away the top 5% of sites that either strongly supported or strongly did not support Tree A. Using this trimmed alignment, RAxML still produced Tree A, indicating that the phylogenetic signal is not driven by a small subset of sites. Finally, I generated maximum likelihood trees using RAxML[332] with an LG4X model for additional GBlocks[343] trimmed alignments consisting of 200, 500, and 3,298 orthologous genes (the latter being all genes with 1-1 orthologs across all 12 species). All alignments resulted in Tree A with good bootstrap support[331].

I generated separate phylogenetic trees using RAxML for each gene in the 1,000 orthologous gene alignment using their optimal substitution models (see above). The consensus tree, as calculated using RAxML[332], of these gene-specific trees was Tree A. Most nodes were supported by the majority of gene trees (>50%). The two nodes that differ between Tree A and Tree B are less well supported. Only 25% of genes support the placement of *P. malariae* as the distant outgroup, while 38% support *P. ovale* branching off with the rodent malaria parasites. While these percentages may seem low, a lower

Table B.1: Effects of Different Filtering Methods on Resulting Phylogenetic Tree

Trimming Method	Retained Amino Acids	Phylogenetic Tree
Untrimmed	1,012,857	Tree C
GBlocks (default param.)	421,988	Tree A
GBlocks (liberal param.)	480,453	Tree C
TrimAl (nogap)	569,568	Tree C
TrimAl (strict)	450,571	Tree A
TrimAl (strictplus)	418,240	Tree A

proportion of genes support Tree B, namely only 19% of genes support the placement of *P. ovale* as an outgroup to *P. malariae* and *P. vivax*, while 24% place *P. malariae* between *P. ovale* and *P. vivax*. This shows that there is significant heterogeneity in the phylogenetic signal present and that these particular nodes are difficult to resolve. I however show that a larger proportion of genes- and the strongest signals when alignments of all genes are concatenated- support Tree A than support Tree B.

B.2 Alignment Effect on Tree Topology

In order to determine the effect of alignment filtering on the resulting tree, I constructed phylogenetic trees using RAxML, with the JTT amino acid substitution model as in Ansari et al. [10], using a number of different alignment trimming strategies. This included no trimming, trimming using GBlocks [343] default parameters (as above), loosening the GBlocks [343] parameters by allowing some gapped sites (if in less than 50% of sequences) and allowing smaller synteny blocks (down to 2 sites), as well as all three TrimAl [49] preset options (nogap, strict, strictplus), as used in Ansari et al. [10] (table B.1).

Of the six trimming methods, the three most stringent filtering methods all resulted in Tree A. The other three trees labeled as 'Tree C' were identical to each other, not placing *P. ovale* as a sister taxon with the rodent malaria parasites but still differing from Tree B by placing *P. malariae* in the same position as Tree A. I was therefore unable to generate Tree B using my alignment. I show however

Table B.2: Correlations of molecular evolution coefficients with number of sites discarded (Absolute) and proportion retained (Retained)

	HKAr	HKAr	Ka/Ks	Ka/Ks	MK (p-value)	MK (p-value)
Comparison	Absolute	Retained	Absolute	Retained	Absolute	Retained
P. o. curtisi/ P. o. wallikeri	0.099*	-0.066	0.324***	-0.478***	-0.179***	0.025
P. malariae/ P. malariae-like	0.139**	-0.088	0.312***	-0.365***	-0.178***	0.090
P. falciparum/ P. reichenowi [259]	0.094	-0.115*	0.454***	-0.554***	-0.194***	0.107*

* p < 0.1, ** p < 0.01, *** p < 0.001

that the stringency of the alignment filtering can severely impact the placement of *P. ovale*. The more stringent the filtering, the more likely *P. ovale* is placed as a sister taxon to the rodent malaria species. Due to the impact that filtering has on determining the topology, I investigated whether the filtering performed using GBlocks [343] with default parameters is appropriate. I correlated (Pearson's correlation coefficient, r) the absolute number of sites removed (Absolute) and the proportion of the gene sequence retained (Retained) with a number of molecular evolution selection coefficients (HKAr, Ka/Ks, MK p-value) calculated for three species-species comparisons (Appendix A) for each gene. Note that these statistics were calculated using variant calls directly from mapped reads, and so should be relatively robust to alignment quality itself. Table B.2 shows the Pearson's correlation coefficient (r) and Bonferroni adjusted p-values for those correlations.

The table shows that there is a strong correlation between Ka/Ks and MK for all three comparisons with the number of sites removed from each gene. The strong positive correlation for Ka/Ks indicates that Ka/Ks is higher in genes where more of the gene sequence is removed. The strong negative correlation for MK (p-value) indicates that the p-values tend to be lower (*i.e.* more likely to be significant) in genes where more of the gene sequence is removed. Hence, genes that seem to be under significant selective pressures tend to be filtered more heavily using GBlocks [343]. This is as expected, as Ka/Ks

measures are increased by bad alignments. This means that the filtered alignment consists of a larger proportion of neutrally evolving sites, which are more informative for phylogenetic inference.

I performed a GO term enrichment analysis[3] by looking at the top 10% of genes that were filtered either the most or the least by GBlocks[343]. I found a very strong enrichment for 'GO:0006412: translation' ($p < 6 \times 10^{-7}$) in the highly filtered genes, in addition to a number of ribosomal GO terms: 'GO:0022625: cytosolic large ribosomal subunit' ($p < 0.001$) and 'GO:0022627: cytosolic small ribosomal subunit' ($p < 0.001$). I did not see enriched GO terms in the genes that were not filtered much. Ribosomal genes are often either extremely conserved or highly variable, making them difficult to align and they were therefore filtered away by GBlocks[343]. Many of the genes that I filtered away using GBlocks[343], including ribosomal genes and surface antigens such as *ama1*, were previously included in a manually selected genelist that generated Tree B[10].

The genes chosen by Ansari et al.[10] are enriched for those with poorer-quality sequence alignments and showing signals of selection. My alignment filtering approach reveals a more robust signal for an alternative topology (Tree A). In any case, I note that even without filtering, a comprehensive analysis of one-to-one orthologs between *Plasmodium* species does not support Tree B, but agrees with our preferred tree except in the precise placement of *P. ovale*.

C

Chapter 2 Methods

C.1 Ethics Statement

The protocol used to collect human blood samples for patients with malaria attending Royal Darwin Hospital was approved by the Health Research Ethics Committee of Menzies School of Health Research (HREC 09/83). Written informed consent was obtained from the patient.

C.2 Sample Collection

Plasmodium malariae DNA used in this study was isolated from a symptomatic patient who presented to the Royal Darwin Hospital, Australia in March and April 2015 with a *P. malariae* parasitemia detected by blood film examination. At each episode, 5ml of EDTA blood was collected from the patient for routine confirmation of malaria by microscopy, full blood count, urea and electrolytes and liver function tests. An additional 10ml of EDTA blood was collected and leukodepleted by passage through a Plasmodipur filter (Euro-diagnostica) within 6 hours of collection. DNA was extracted from a 2ml aliquot of the filtered red blood cell pellet using the QIAamp DNA Blood Midi Kit (Qiagen) as per the manufacturer's instructions, and stored at ≈20°C. *Plasmodium* species was confirmed by PCR for *P. vivax*, *P. falciparum*, *P. malariae* and *P. ovale* parasites using a modified version of that described by Padley et al. [265] so that each species was identified in a separate (non-multiplex) assay. PCR for *P. knowlesi* parasites was undertaken using the method of Imwong et al. [151].

C.3 Genome Sequencing

Whole genome sequencing was performed on both parasite isolates (PmUG01 and PmUG02) using Illumina Standard libraries of 200–300bp fragments and amplification-free libraries of 400–600bp fragments were prepared[43] and sequenced on the Illumina HiSeq 2000 v4, the MiSeq v2, and the X Ten according to the manufacturer's standard protocol. Raw sequence data were deposited in the European Nucleotide Archive (table C.1).

Table C.1: Sequencing Information and Statistics

Characteristic	PmUG01	PmUG02
Accession Number	ERS1110316	ERS1110319
Origin	Uganda	Uganda
Infection	Initial	Recrudescence
Mean Coverage Depth	407x	136x
Coverage Range (Min-Max)	0x-7499x	0x-3519x
% Genome Covered at 1x	99.9%	99.9%
Sequencing Platform	Illumina HiSeq X Ten, Illumina MiSeq v2	Illumina HiSeq X Ten, Illumina MiSeq v2
Library Type	Amplification Free	Amplification Free

C.4 Genotyping of Single Nucleotide Variants

The two *P. malariae* samples (PmUG01, PmUG02) were mapped against the *P. malariae* reference genome[301] using SMALT (-y 0.8, -i 500). The resulting bam format files were merged, and GATK's[215] UnifiedGenotyper was used to call SNPs from the merged bam format files (table C.2). According to GATK's[215] best practices, SNPs were filtered by quality of depth (QD >2), depth of coverage (DP >20), mapping quality (MQ >30), and strand bias (FS <60). SNPs in low-complexity regions, as determined by DUSTmasker[227], were removed, as were sites with missing data in either of the two samples, and SNPs within 50bp of each other to avoid SNPs in repetitive regions. Finally, only exonic SNPs were retained. We performed the same SNP calling procedure by also including additional previously published samples[301]. Heterozygous sites were filtered out, while SNPs in non-coding regions were retained (table C.3). Raw SNPs for PmUG01 and PmUG02 differ between table C.2 and table C.3 due to calling SNPs from a merged bam format file instead of individually. Samples are therefore pooled and SNP calling is performed on the population rather than on the individual. The consequence of this is that if an individual sample has insufficient reads at a particular locus to reliably call

Table C.2: SNP Calling Results specifically for both PmUG01 and PmUG02

Sample ID	PmUG01	PmUG02
Raw SNPs	274,494	287,610
Private	79,953	93,069
Ref	362,848	333,615
Missing*	298	1,470
Filtered SNPs	2,442	1,499
Private	1,132	189
Ref	189	1,773
Missing*	0	0

*Sites at which the sample has no coverage. SNP calling results as per mapping the two *P. malariae* samples from the present clinical case against the PmUG01 reference genome [301]. The raw SNPs are the total number of SNPs that we called using GATK's UnifiedGenotyper default parameters in the different samples. Of these raw SNPs, some are exclusive to a certain sample (Private), are identical to the reference genome (Ref), or there is no coverage and therefore no SNP call could be made (Missing). The same information is also shown for the filtered SNPs, which were filtered according to several different parameters.

a SNP there, that SNP might still be called if other samples in the population also have SNPs at that location, as this increases the likelihood of a SNP at that position. The number of SNPs therefore differs for identical samples depending on the population on which the SNP calling was performed.

C.5 Abundance Calculations

Using the relative SNP frequencies of the three haplotypes (R1, H1 and H2), the relative abundances of the different haplotypes were calculated, assuming that the number of sequencing reads is proportional to the abundance of the specific haplotype in the blood. Manually inspecting the SNP frequencies by eye, H1 and R1 are approximately in a ratio of 35:65 and H2 to R1 is in a ratio of 15:85. In both cases I ignore the third haplotype because I cannot ascertain its genotype. Ratio multiplication yields a joint ratio of 975:2975:5525 for H2:H1:R1, simplifying to approximately 10:30:60. Explicitly, multiplying 85 by 65 (the R1 terms in both ratios) gives 5525 for the above joint ratio. We know that H1 is in a 35:65 ratio to R1, so the joint ratio term for H1 is 5525 x (35/65) resulting in 2975, and

Table C.3: SNP Calling Results for all P. malariae samples

Sample ID	PmUG01	PmUG02	PmMY01	PmID01	PmMA01	PmGN01
Origin	Uganda	Uganda	Malaysia	Papua Indonesia	Mali	Guinea
Raw SNPs	200,679	191,766	252,172	187,327	198,029	503,175
Private	13,300	10,462	48,632	19,396	26,637	73,208
Ref	531,180	479,142	367,672	347,257	360,155	398,393
Missing*	5,468	23,064	65,361	107,355	95,973	38,609
Filtered SNPs	1,375	2,707	22,696	17,564	16,057	21,329
Private	0	414	8,816	5,939	6,079	10,343
Ref	49,647	47,868	27,521	32,762	34,205	29,021
Missing*	0	0	0	0	0	0

*Sites at which the sample has no coverage. SNP calling results as per mapping all *P. malariae* samples against the PmUG01 reference genome [301]. The raw SNPs are the total number of SNPs that we called using GATK's UnifiedGenotyper default parameters in the different samples. Of these raw SNPs, some are exclusive to a certain sample (Private), are identical to the reference genome (Ref), or there is no coverage and therefore no SNP call could be made (Missing). The same information is also shown for the filtered SNPs, which were filtered according to several different parameters.

the same logic holds for H2 giving 975. The joint ratio is a way of representing the relative abundances of the haplotypes but do not inform us on their absolute abundances in the infection.

Tri-allelic sites offer the most straightforward way of observing the ratio of the three haplotypes, however their number is low. Following SNP filtering (see above), the retained 13 tri-allelic sites were spread evenly across the genome (table C.4). Assuming that the allele with the highest depth is R1, the intermediate depth is H1, and lowest depth is H2, I calculated the mean depth for all three. This yielded a ratio of ≈9:22:69 for H2:H1:R1 (table C.4). The ratio for PmUG02 was also calculated (table C.5).

Table C.4: Sequencing reads for the three haplotypes in tri-allelic sites in PmUG01

Genomic Location	H2 haplotype reads	H1 haplotype reads	R1 haplotype reads
Chr1: 704,003	6	21	202
Chr2: 437,291	43	71	130
Chr7: 1,580,127	12	18	75
Chr9: 1,429,882	21	82	140
Chr10: 873,165	9	31	209
Chr11: 987,107	17	60	173
Chr11: 2,535,419	47	84	113
Chr12: 2,746,074	9	44	192
Chr12: 2,855,999	11	36	176
Chr12: 3,047,435	11	34	204
Chr13: 1,061,473	13	16	221
Chr14: 1,118,313	27	62	139
Total (Proportion)	265 (0.09)	650 (0.22)	2,091 (0.69)

Assuming that the lower-level genotype is H2, intermediate-level genotype is H1, and higher-level genotype is R1, the tri-allelic sites have a certain number of sequencing reads confirming each of these three genotypes.

Table C.5: Sequencing reads for the three haplotypes in tri-allelic sites in PmUG02

Genomic Location	H2 haplotype reads	H1 haplotype reads	R1 haplotype reads
Chr1: 704,003	16	0	7
Chr2: 437,291	20	4	0
Chr7: 1,580,127	18	0	9
Chr9: 1,429,882	25	0	0
Chr10: 873,165	23	0	0
Chr11: 987,107	19	8	1
Chr11: 2,535,419	11	0	7
Chr12: 2,746,074	15	0	1
Chr12: 2,855,999	9	0	1
Chr12: 3,047,435	6	4	4
Chr13: 1,061,473	8	0	5
Chr14: 1,118,313	26	0	0
Total (Proportion)	212 (0.81)	16 (0.06)	35 (0.13)

Haplotype assignment based on sequencing read ratios from the PmUG01 sample.

D

Chapter 3 Methods

D.1 Sample Collection and Preparation

From 2010 to 2013, patients with uncomplicated *P. falciparum* malaria were enrolled in a parasite clearance rate studies[6,7,15] in three provinces in Cambodia. The three provinces were selected based on the level of artemisinin and piperaquine resistance in the local parasite population: Pursat (common), Preah Vihear (emerging), and Ratanakiri (uncommon)[8]. Informed consent was obtained in writing from adult patients or by a guardian in the case of child patients. Protocols were approved

by the Cambodian National Ethics Committee for Health Research and the National Institute of Allergy and Infectious Diseases Institutional Review Boards. The protocols are registered with ClinicalTrials.gov, under the numbers NCT00341003, NCT01240603, and NCT01736319.

Whole blood samples were obtained from patients at enrolment. The samples were then leukocyte-depleted using the CF11 filtration method[355], followed by DNA extraction using the QIAamp DNA Blood Kit (Qiagen, Valencia, CA). Relative ratios of human to *Plasmodium* DNA were ascertained using a Qubit instrument (Invitrogen, Carlsbad, CA) based fluorescence analysis, as well as through the use of a multi-species quantitative PCR run on the Roche Lightcycler 480 II system[209].

D.2 *In-vitro* Drug Assays

The *in-vitro* drug assays using a standard 72hr SYBR Green I-based stain-based method[209] were performed for chloroquine and mefloquine on parasites freshly obtained from patients. The IVART software[198] was used to determine IC_{50} values by fitting the drug concentration-growth inhibition data. Throughout these assays, the *P. falciparum* 3D7 line was used as a quality control.

D.3 Whole Genome Sequencing

Samples with low levels of human contamination (<80% human) and sufficient DNA (>50ng) were selected for whole genome sequencing. Sequencing was performed on the Illumina HiSeq platform following the manufacturer's standard protocols[28], producing 100bp paired-end sequencing reads. Approximately 1Gbp of read data was generated per sample.

D.4 SNP Calling and Filtering

Single nucleotide polymorphisms (SNPs) were typed across the *P. falciparum* genome in positions included in the high-quality SNP set established via the MalariaGEN community project V6. SNPs were extracted for samples with mefloquine IC_{50} values and filtered as follows. In addition to the filtering criteria defined in the MalariaGEN community project V6 (flag: "PASS"), the following filtering criteria were used to remove alignment artefacts and low-confidence genotype calls: insertions/deletions; non-coding SNPs; SNPs with >2 alleles; SNPs with >20% missingness; SNPs with minor allele frequency less than 3%; SNPs in plastids. Prior to the filtering, heterozygous calls were manually recoded as missing genotype calls.

D.5 Genome-Wide Association Study (GWAS)

The GWAS analyses were performed using a linear mixed model implemented in FaST-LMM [199] version 2.06. I performed the GWAS with the log base 10 transformed mefloquine IC_{50} values as the continuous dependent variable. I corrected for population structure using a relationship matrix [221], calculated from a subset of unlinked SNPs (in windows of 100 SNPs, shifted forward by 10 SNPs each time, removing one from each pair of SNPs with linkage disequilibrium >0.3) using PLINK version 2 [56,287] (option: –indep-pairwise 100 10 0.3). I also always included the province of origin for each sample as a covariate (encoded as 1, 2 and 3 for Pursat, Preah Vihear, and Ratanakiri, respectively). To correct for multiple testing, I corrected the p-value threshold for genome-wide significance using the Bonferroni method, yielding a threshold of $p < 4 \times 10^{-6}$ for both phenotype datasets. I also utilized a second, less stringent, suggestive threshold of $p < 1 \times 10^{-4}$ to look at other SNPs highly associated

with the phenotype. Depending on the GWAS, additional covariates were included, CNV or F1068L *mdr1* SNP presence/absence, in each case, I encoded the wild type as '1', the alternate as '2' and missing data as '0'.

D.6 Copy Number Amplification Calling

Copy number amplification calling was performed by Richard Pearson from the Wellcome Sanger Institute. Two orthogonal methods were used to determine presence/absence of *mdr1* and *plasmepsin 2/3* CNVs: a coverage-based method and a method based on position and orientation of reads near discovered duplication breakpoints. Briefly, a coverage-based hidden Markov model was used to identify potential copy number amplifications and their boundaries. Breakpoints of duplications around *mdr1* and *plasmepsin 2/3* were manually inspected by identifying face-away read pairs spanning the supposed breakpoints. If face-away read pairs confirmed the prediction of the hidden Markov coverage-based model, then a CNV was called as present. If no coverage was present in the breakpoint region, then the CNV genotype was called 'missing'.

D.7 Miscellaneous

All the SNP filtering and GWAS analyses were run in the Juptyer notebook suite [169]. Downstream analysis was performed using R version 3.3.2 within RStudio [271]. The beeswarm R package was used for plotting the phenotype distributions. Figures were arranged using the Inkscape software.

E

Chloroquine GWAS

E.1 Spatial & Geographical Trends in Chloroquine Resistance

Originating from three Cambodian provinces (Pursat, Preah Vihear, and Ratanakiri), 391 clinical *P. falciparum* isolates collected between 2010 and 2013 were phenotyped for chloroquine (CQ) 50% inhibitory concentration (IC_{50}) (table E.1). I observed significant differences in CQ IC_{50} between the different provinces, with Pursat having a significantly higher mean CQ IC_{50} of 456 nmol/L compared to 355 nmol/L in Preah Vihear (t-test: $p < 0.005$) and 213 nmol/L in Ratanakiri ($p < 5 \times 10^{-16}$) (figure

Table E.1: Sample Origin Information

Year	Pursat	Preah Vihear	Ratanakiri	Total
2010	48	0	39	87
2011	79	59	49	187
2012	22	33	22	77
2013	14	14	12	40
Total	163	106	122	391

E.1 a). The mean CQ IC_{50} value in Preah Vihear is also significantly higher than in Ratanakiri ($p < 5 \times 10^{-9}$). I also noticed significant increases in CQ IC_{50} in all provinces from the start of collection to the end of the collection (figure E.1 b). In Pursat, the mean CQ IC_{50} increased from 447 nmol/L in 2010 to 703 nmol/L in 2013 (t-test: $p < 0.05$). In Preah Vihear, it increased from 294 nmol/L in 2011 to 621 nmol/L in 2013 ($p < 0.0005$), while in Ratanakiri it went from 202 nmol/L in 2010 to 409 nmol/L in 2013 ($p < 0.05$).

E.2 GWAS of Chloroquine Resistance

To better understand the genetic basis underpinning the differences in CQ levels, whole genome sequencing was performed on all samples with CQ IC_{50} values. In order to identify genetic markers that may be associated with the different levels of CQ and MQ resistance, I performed a GWAS for the 391 *P. falciparum* samples with CQ IC_{50} values.

Beginning with the samples with CQ IC_{50} values, following a variety of filtering steps (Appendix D), I identified 12,603 high confidence SNPs in coding regions that had a non-reference genotype call in at least 12 samples (3% of all samples). I associated these 12,603 SNPs with CQ $\log(IC_{50})$ values as a continuous dependent variable using the fastLMM software [199], which employs a linear mixed model algorithm. I controlled for confounding effects by treating the province of origin as a covariate and

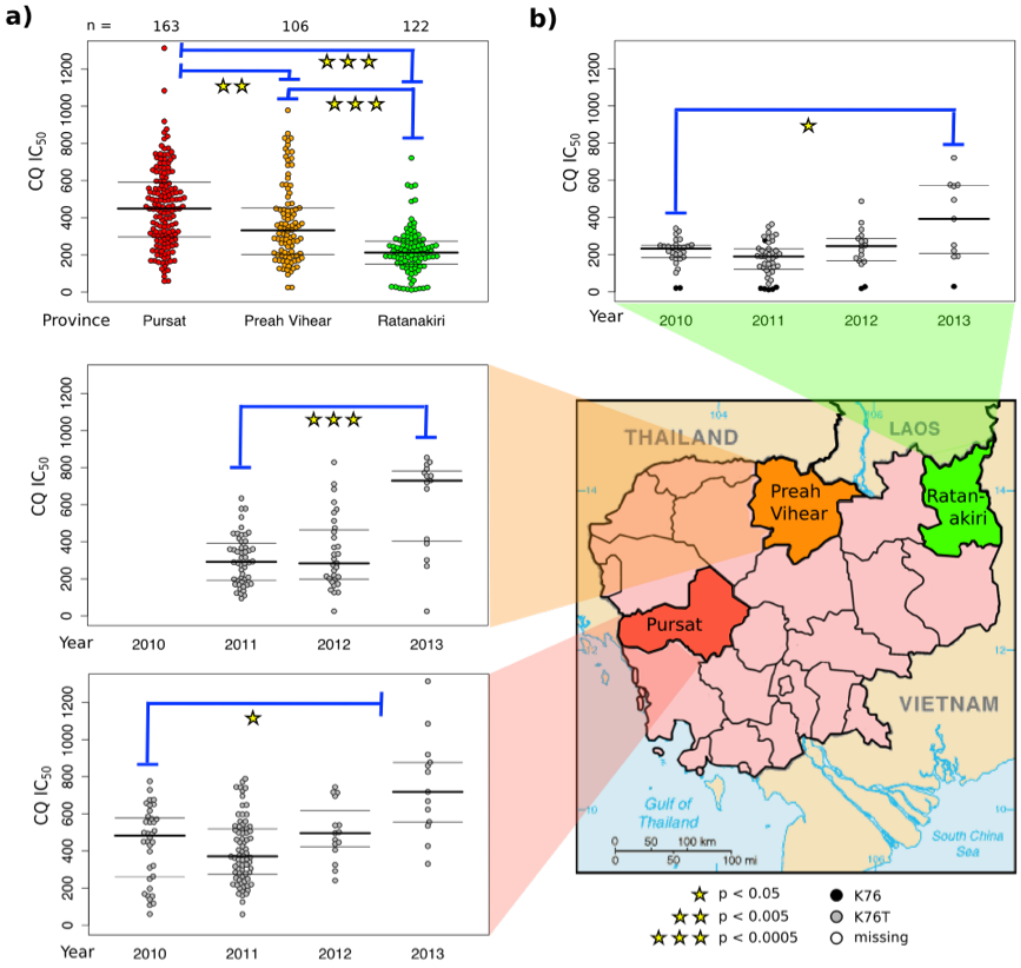


Figure E.1: Differences in chloroquine CQ IC_{50} between different provinces in Cambodia (a) and across the years of collection within each province (b). Each point represents one clinical P. falciparum isolate, colored either by province of origin (a) (see map) or by pfcrt K76 genotype (b) (see legend). The number of samples per province is indicated above (a). The map shows the respective geographical location of the three provinces within Cambodia. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (see legend).

population structure (as measured by genetic similarity across samples) as a random effect.

The GWAS yielded a very strong signal on Chromosome 7, with all the SNPs that passed the genome-wide threshold for significance (Bonferroni-adjusted p-value < 4×10^{-6}) lying within the *pfcrt* gene (PF3D7_0709000) (figure E.2 a)(table E.2). The SNP most strongly associated with CQ log(IC_{50}) is the K76T mutation, currently used as the most sensitive marker of CQ resistance [81]. Two other significant SNPs (Q271E and A220S) are known to be commonly associated with the K76T mutation [103,90] and are thought to be required in the mutational pathway to achieve chloroquine resistance [337]. I observed a significant difference in CQ IC_{50} between K76 wild type and mutant parasites (t-test: $p < 1.4 \times 10^{-9}$) (figure E.2 b), but still note a striking level of variance in the CQ IC_{50} values of K76T carrying parasites (figure E.2 b). Indeed, the differences in the level of CQ resistance between the three provinces and across time seem to be independent of the K76T mutation (figure E.1 b).

The second and third most significant SNPs in the GWAS analysis code for amino acid changes just upstream of the K76T mutation (N75D, M74I) (figure E.2 a & table E.2). These SNPs encode a haplotype unique to Cambodia known as CVIDT (from amino acid positions 72-76) [197], distinguishing it from the wild type CVMNK haplotype. The most common haplotype besides CVMNK is the CVIET haplotype [375], which is the CQ resistant haplotype that initially spread from Southeast Asia to Africa [12]. The CVIET haplotype is formed by multiple mutations in the 74 and 75 codon positions that are called as an indel using automated genotype callers, resulting in missing genotype calls for the SNPs at these positions due to filtering out indels. Encoding this CVIET haplotype as a biallelic SNP, I find that it also passes the suggestive threshold I employed (figure E.2 a & table E.2). Comparing the CQ IC_{50} values of CVIDT and CVIET samples, I find a significantly higher level of CQ IC_{50} in CVIET samples compared to CVIDT ones (figure E.2 c), as has previously been suggested [90]. Addi-

Table E.2: SNPs most strongly associated with CQ log(IC_{50})

Chromosome	Position	Gene ID	Gene Description	N or S	Alteration	p-value
7	403,625	PF3D7_0709000	Chloroquine resistance transporter	N	K76T	1.8×10^{-20}
7	403,621	PF3D7_0709000	Chloroquine resistance transporter	N	N75D	2.0×10^{-17}
7	403,620	PF3D7_0709000	Chloroquine resistance transporter	N	M74I	2.0×10^{-17}
7	404,836	PF3D7_0709000	Chloroquine resistance transporter	N	Q271E	5.3×10^{-17}
7	404,407	PF3D7_0709000	Chloroquine resistance transporter	N	A220S	9.4×10^{-17}
7	412,466	PF3D7_0709200	Glutaredoxin-like protein	S	n/a	2.5×10^{-5}
7	≈403,620	PF3D7_0709000	Chloroquine resistance transporter	Indel	MN74.75IE	6.4×10^{-5}
7	410,036	PF3D7_0709100	Cg1 protein	N	N608D	9.0×10^{-5}

SNPs that pass either the Bonferroni-adjusted p-value threshold ($p < 4 \times 10^{-6}$) or the more lenient suggestive threshold ($p < 1 \times 10^{-4}$) are listed in order of increasing p-value. The table shows chromosome and nucleotide position of the SNP, the ID and description of the gene in which the SNP occurs, whether it is a synonymous (S) or nonsynonymous (N) mutation, what amino acid alteration it encodes if it is nonsynonymous, and the p-value associated with the SNP.

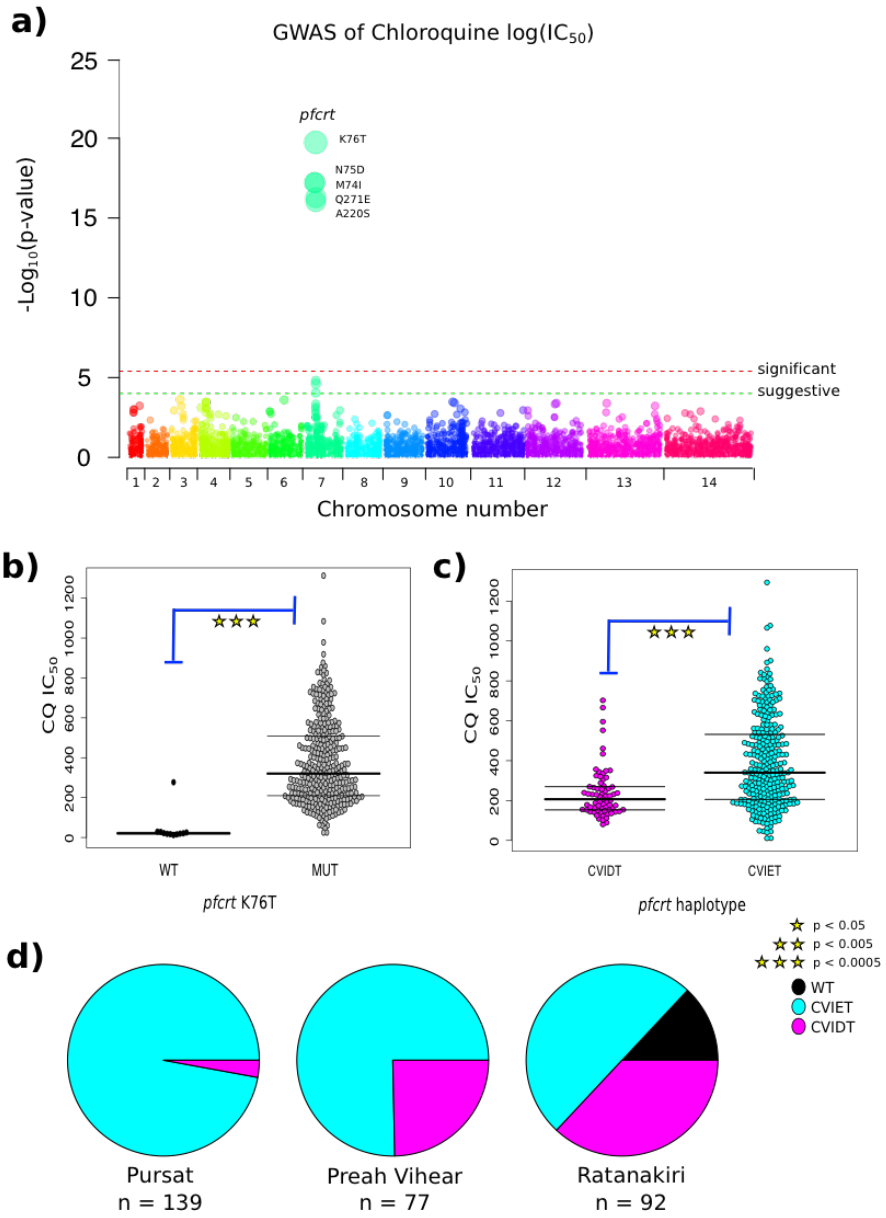


Figure E.2: a) GWAS analysis using chloroquine log(IC_{50}) as the dependent variable. Each point corresponds to a SNP, coloured by the chromosome it is located on and they are ordered by their position on the chromosome. The dotted red line indicates the genome-wide threshold for significance (4×10^{-6}) and the dotted green line is a more lenient suggestive threshold (1×10^{-4}). The SNPs exceeding the significance threshold are labeled by the gene they are found in and the amino acid alteration they code for. b) Difference in CQ IC_{50} between parasites without (black) and with (grey) the K76T mutation in pfcrt. c) Difference in CQ IC_{50} between parasites with a CVIDT (magenta) and a CVIET (blue) haplotype for pfcrt. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (see legend). d) Pie charts of the differential distribution of CVIET (blue) and CVIDT (magenta) pfcrt haplotypes between the three different provinces. Number of samples shown below the pie charts.

tionally, the relative ratio of these two haplotypes to each other differs substantially between the three sampled provinces (figure E.2 d), with Pursat having an overrepresentation of CVIET parasites and Ratanakiri having a comparatively high proportion of CVIDT samples. This therefore mirrors the overall differences in chloroquine resistance between the different provinces.

E.3 Additional GWAS Analysis Identifies Novel Marker

Following the observation that only SNPs within *pfcrt* passed the genome-wide significance threshold, I wanted to investigate whether there are any novel markers of chloroquine resistance in other parts of the genome. I redid the GWAS using the K76T SNP genotype as an additional covariate. Doing this however, the signal obtained in the original GWAS disappears, with no SNP passing the genome-wide significance threshold (figure E.3 a).

Differences in high-end CQ IC_{50} values being more pronounced on a linear scale compared to a log scale, I performed a GWAS with CQ IC_{50} as the dependent variable, instead of CQ $\log(IC_{50})$. Interestingly, I do not observe a significant SNP in the *pfcrt* gene, showing that the significance of the K76T mutation is visible only when looking on a log-transformed scale (table E.3 & figure E.3 b). I identified three potential SNPs of interest, one of which passes the genome-wide threshold for significance. It is a nonsynonymous SNP (D2658Y) in a *dynein heavy chain protein* gene (PF3D7_1202300). Samples with this particular mutation and K76T seem to have a significantly higher CQ IC_{50} than samples that only have K76T (figure E.4 a). It is found in samples from all three Cambodian provinces, but has increased dramatically in frequency across the four years of samples, with almost a quarter of samples possessing this SNP in 2013 (figure E.4 b). This doesn't seem to be due to a clonal expansion, as the 19 samples with this mutation do not seem to be related genetically (figure E.4 c), and the flanking

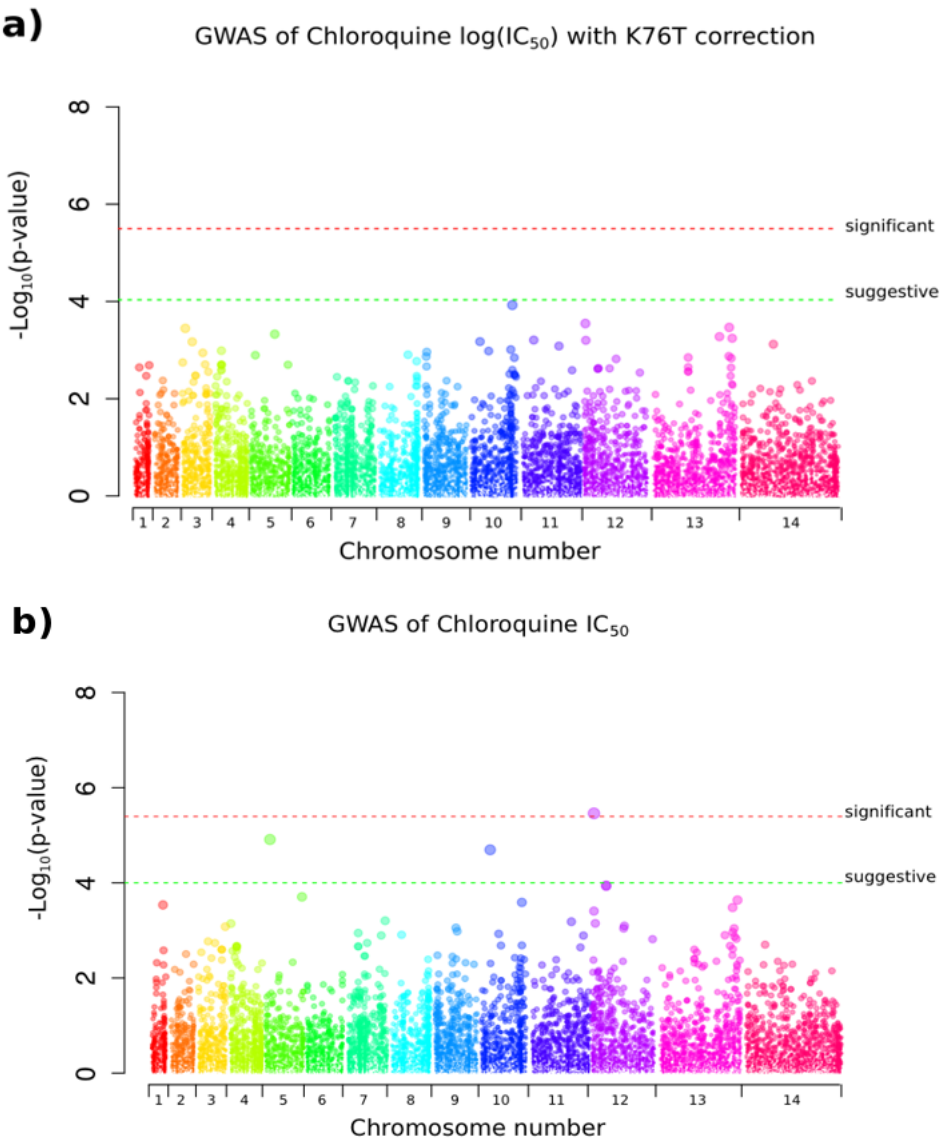


Figure E.3: GWAS analyses using (a) the K76T genotype in pfcrt as a covariate and (b) the linear CQ IC_{50} as the dependent variable. In both analyses, I controlled for population structure and sample origin. Each point corresponds to a SNP, coloured by the chromosome it is located on and they are ordered by their position on the chromosome. The dotted red line indicates the genome-wide threshold for significance ($p < 4 \times 10^{-6}$) and the dotted green line is a more lenient suggestive threshold ($p < 1 \times 10^{-4}$).

Table E.3: SNPs most strongly associated with linear CQ IC_{50}

Chromosome	Position	Gene ID	Gene Description	N or S	Alteration	p-value
12	129,899	PF3D7_1202300	Dynein heavy chain	N	D2658Y	3.5×10^{-6}
5	202,752	PF3D7_0504800	Conserved Plasmodium protein	N	G661R	1.2×10^{-5}
10	372,716	PF3D7_1009100	Conserved membrane protein	N	L66F	2.0×10^{-5}

Table E.4: SNPs that pass either the Bonferroni-adjusted p-value threshold ($p < 4 \times 10^{-6}$) or the more lenient suggestive threshold ($p < 1 \times 10^{-4}$) are listed in order of increasing p-value. The table shows chromosome and nucleotide position of the SNP, the ID and description of the gene in which the SNP occurs, whether it is a synonymous (S) or nonsynonymous (N) mutation, what amino acid alteration it encodes if it is nonsynonymous, and the p-value associated with the SNP.

regions around the SNP are quite diverged between the samples (figure E.4 d). This therefore seems like an interesting candidate SNP to look further into, though it is unclear how a SNP in a *dynein heavy chain protein* gene would increase a parasite's CQ resistance from a functional perspective.

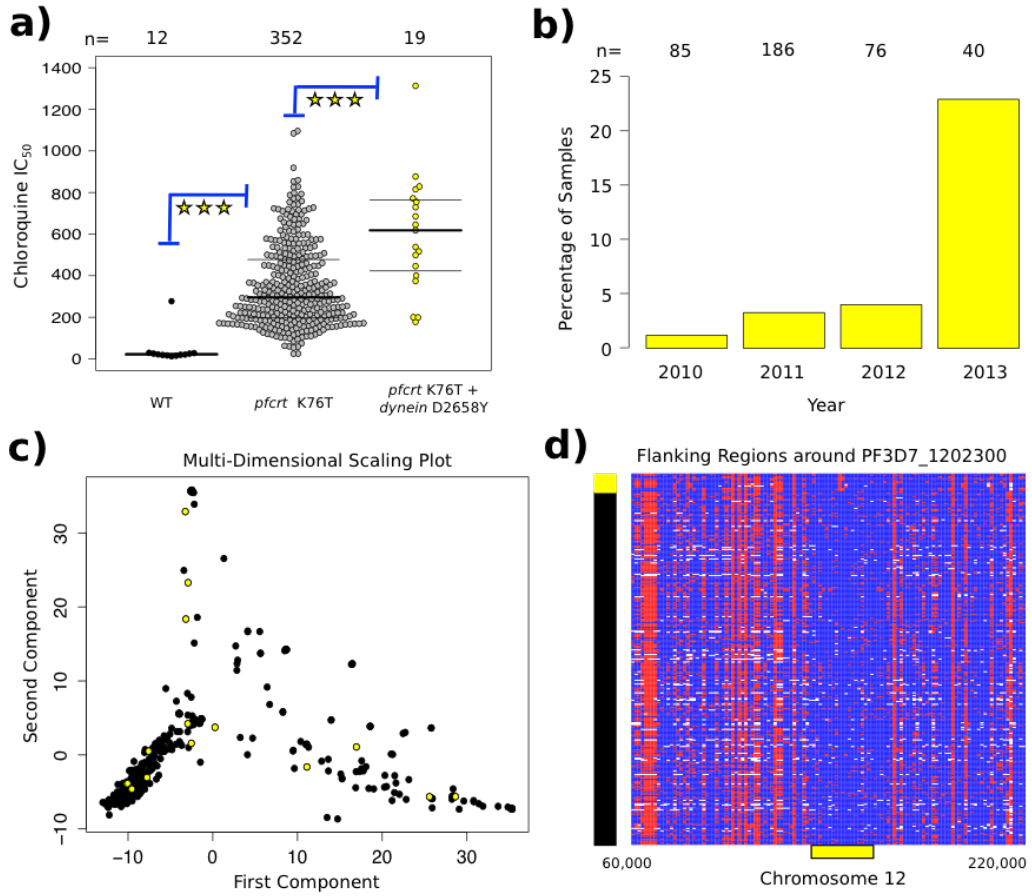


Figure E.4: a) Difference in CQ IC_{50} between wild type parasites (black), those with the K76T mutation in pfcrt (grey) and those with both the K76T mutation and the D2658Y SNP in the dynein heavy chain gene (PF3D7_1202300) (yellow). Numbers above plot indicate sample numbers. The horizontal lines indicate the median (bold) and interquartile range (thin) of the respective distributions. Blue bars indicate Welch two-sample t-tests performed between two distributions, with stars indicating the level of significance of that comparison (*** = $p < 0.0005$). b) Percentage of samples with the D2658Y SNP in PF3D7_1202300 across the four years of sampling. Numbers above the plot indicate the number of samples. c) A multidimensional scaling plot for all 391 samples with CQ IC_{50} values, based on all 12,603 SNPs called genome-wide. Yellow dots represent samples with the D2658Y mutation. d) Flanking regions around PF3D7_1202300, showing the genotype of the 391 samples (rows) for all SNPs within 100kb of either side of PF3D7_1202300. The genotypes are either reference (blue), alternate (red), or missing (white). The position of PF3D7_1202300 is indicated with a yellow box. The bar to the left shows which samples have the D2658Y mutation (yellow) and which don't (black).

F

Chapter 4 Methods

F.1 Data and Filtering

The MalariaGEN community project V6 (Pf6) dataset was harnessed for this analysis. Single nucleotide polymorphisms (SNPs) were typed across the *P. falciparum* genome in positions included in the high-quality SNP set established via the MalariaGEN community project V6. In addition to the filtering criteria defined in the MalariaGEN community project V6 (flag: 'PASS'), the following filtering criteria were used to remove alignment artefacts and low-confidence genotype calls: inser-

tions/deletions; non-coding SNPs; SNPs with >2 alleles; SNPs with >20% missingness; SNPs with minor allele frequency less than 3%; SNPs in plastids. Prior to the filtering, heterozygous calls were manually recoded as missing genotype calls.

F.2 Antimalarial Drug Resistance Phenotype Data

The *in-vitro* drug assays using a standard 72hr SYBR Green I-based stain-based method [209] were performed for chloroquine, piperaquine, and mefloquine on parasites freshly obtained from patients. The IVART software [198] was used to determine IC_{50} values by fitting the drug concentration-growth inhibition data. Throughout these assays, the *P. falciparum* 3D7 line was used as a quality control. The artemisinin parasite clearance half-life ($PC_{t1/2}$) phenotype data was obtained from the TRAC study (NCT01350856) and the US National Institutes of Health study (NCT00341003 and NCT01240603). In brief, during treatment, parasite densities were estimated by counting parasitized erythrocytes in blood smears from peripheral blood samples taken at 0, 4, 6, 8 and 12 h after patient admission and then every 6h until two consecutive counts were negative. $PC_{t1/2}$ estimates were computed from these parasite counts, by fitting a statistical model [104] using the Parasite Clearance Estimator developed by WWARN.

F.3 Haplogroup Classification

Haplogroup classification was performed by Jacob Almagro Garcia from the Big Data Institute in Oxford. Both *pfcrt* and *mdr1* genes were classified into haplogroups using the filtered Pf6 MalariaGEN data set (www.malariagen.net). Samples with a missing genotype call in either gene were not classified for that gene respectively. Haplogroups were then determined using all nonsynonymous mutations

in the coding regions of the genes. Thus, genes differing by one or more nonsynonymous mutations were categorized into separate haplogroups.

F.4 Copy Number Variation Calling

Copy number amplification calling was performed by Richard Pearson from the Wellcome Sanger Institute. Two orthogonal methods were used to determine presence/absence of *mdr1* and *plasmepsin 2/3* CNVs: a coverage-based method and a method based on position and orientation of reads near discovered duplication breakpoints. Briefly, a coverage-based hidden Markov model was used to identify potential copy number amplifications and their boundaries. Breakpoints of duplications around *mdr1* and *plasmepsin 2/3* were manually inspected by identifying face-away read pairs spanning the supposed breakpoints. If face-away read pairs confirmed the prediction of the hidden Markov coverage-based model, then a CNV was called as present. If no coverage was present in the breakpoint region, then the CNV genotype was called 'missing'.

F.5 Miscellaneous

All the SNP filtering was performed in the Juptyer notebook suite [126]. Downstream analysis was performed using R version 3.3.2 within RStudio [271]. The beeswarm R package was used for plotting the phenotype distributions. Figures were arranged using the Inkscape software.

G

Thesis Outputs

The research that was performed as part of this PhD project has been presented at numerous international conferences and has resulted in a number of scientific publications, including some that are yet to be published. Specifically, the results from Chapter 1 have been published as:

1. **G. G. Rutledge**, U. Böhme, M. Sanders, A. J. Reid, J. A. Cotton, O. Maiga-Ascofare, A. A. Djimde, T. O. Apinjoh, L. Amenga-Etego, M. Manske, J. W. Barnwell, F. Renaud, B. Ollomo, F. Prugnolle, N. M. Anstey, S. Auburn, R. N. Price, J. S. McCarthy, D. P. Kwiatkowski, C. I. Newbold, M. Berriman and T. D. Otto (2017) "*Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution." Nature 542(7639):101-104. doi:10.1038/nature21038

The results presented in Chapter 2 have been published as:

2. **G. G. Rutledge***, I. Marr*, G. K. L. Huang, S. Auburn, J. Marfurt, M. Sanders, N. J. White, M. Berriman, C. I. Newbold, N. M. Anstey, T. D. Otto and R N. Price (2017) "Genomic charac-

terization of recrudescent *Plasmodium malariae* after treatment with artemether/lumefantrine." Emerg Infect Dis 15(23):8. doi: 10.3201/eid2308.161582 *contributed equally

In addition to the above, additional work not directly related to the results presented in this thesis have resulted in a number of additional publications:

3. A. Gilabert, T.D. Otto, **G. G. Rutledge**, B. Franzon, B. Ollomo, C. Arnathau, P. Durand, N.D. Moukodoum, A.P. Okouga, B. Ngoubangoye, B. Makanga, L. Boundenga, C. Paupy, F. Renaud, F. Prugnolle and V. Rougeron (2018) "*Plasmodium vivax-like* genome sequences shed new insights into *Plasmodium* biology and evolution." PLoS Biol 16(8):e2006035. doi: 10.1371/journal.pbio.2006035

4. **G. G. Rutledge** and R. Amato (2018) "Juggling resistance mutations." Nat Rev Microbiol 16: 332. doi:10.1038/s41579-018-0008-1

5. E. Bushell, A. R. Gomes, T. Sanderson, B. Anar, G. Girling, C. Herd, T. Metcalf, K. Modrzynska, F. Schwach, R. E. Martin, M. W. Mather, G. I. McFadden, L. Parts, **G. G. Rutledge**, A. B. Vaidya, K. Wengelnik, J. C. Rayner and O. Billker (2017) "Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes." Cell 170(2):260-272. doi:10.1016/j.cell.2017.06.030

6. E. M. Pasini, U. Böhme , **G. G. Rutledge**, A. Voorberg-Van der Wel, M. Sanders, M. Berriman and T. D. Otto (2017) "An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion." Wellcome Open Res 2:42. doi:10.12688/wellcomeopenres.11864.1

7. **G. G. Rutledge** and C. V. Ariani (2017) "Finding the needle in the haystack." Nat Rev Microbiol 15(3):136. doi:10.1038/nrmicro.2017.7

8. **G. G. Rutledge** and T. D. Otto (2017) "Last parasite standing." Nat Rev Microbiol 15(1): 4. doi:10.1038/nrmicro.2016.181

9. Oyola, S. O., C. V. Ariani, W. L. Hamilton, M. Kekre, L. N. Amenga-Etego, A. Ghansah, **G. G. Rutledge**, S. Redmond, M. Manske, D. Jyothi, C. G. Jacob, T. D. Otto, K. Rockett, C. I. Newbold, M. Berriman and D. P. Kwiatkowski (2016) "Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification." Malar J 15(1): 597. doi:10.1186/s12936-016-1641-7

References

[1] Abeyasinghe, R. R., Galappaththy, G. N., Smith Gueye, C., Kahn, J. G., & Feachem, R. G. (2012). Malaria control and elimination in Sri Lanka: documenting progress and success factors in a conflict setting. *PLoS One*, 7(8), e43162.

[2] Alano, P. (2014). The sound of sexual commitment breaks the silencing of malaria parasites. *Trends in Parasitology*, 30(11), 509–510.

[3] Alexa, A.; Rahnenfuhrer, J. (2016). topGO: Enrichment analysis for gene ontology. *R package version 2.24.0*.

[4] Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malarial infection. *British Medical Journal*, 1(4857), 290–294.

[5] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17), 3389–402.

[6] Amaratunga, C., Lim, P., Suon, S., Sreng, S., Mao, S., Sopha, C., Sam, B., Dek, D., Try, V., Amato, R., Blessborn, D., Song, L. J., Tullo, G. S., Fay, M. P., Anderson, J. M., Tarning, J., & Fairhurst, R. M. (2016). Dihydroartemisinin-piperaquine resistance in *Plasmodium falciparum* malaria in Cambodia: a multisite prospective cohort study. *Lancet Infectious Diseases*, 16(3), 357–365.

[7] Amaratunga, C., Sreng, S., Suon, S., Phelps, E. S., Stepniewska, K., Lim, P., Zhou, C. J., Mao, S., Anderson, J. M., Lindegardh, N., Jiang, H. Y., Song, J. P., Su, X. Z., White, N. J., Dondorp, A. M., Anderson, T. J. C., Fay, M. P., Mu, J. B., Duong, S., & Fairhurst, R. M. (2012). Artemisinin-resistant *Plasmodium falciparum* in Pursat province, western Cambodia: a parasite clearance rate study. *Lancet Infectious Diseases*, 12(11), 851–858.

[8] Amato, R., Lim, P., Miotto, O., Amaratunga, C., Dek, D., Pearson, R. D., Almagro-Garcia, J., Neal, A. T., Sreng, S., Suon, S., Drury, E., Jyothi, D., Stalker, J., Kwiatkowski, D. P., & Fairhurst, R. M. (2017). Genetic markers associated with dihydroartemisinin-piperaquine failure in Plasmodium falciparum malaria in Cambodia: a genotype-phenotype association study. *Lancet Infectious Diseases*, 17(2), 164–173.

[9] Amato, R., Pearson, R. D., Almagro-Garcia, J., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Drury, E., Stalker, J., Miotto, O., Fairhurst, R. M., & Kwiatkowski, D. P. (2018). Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study. *Lancet Infect Dis*, 18(3), 337–345.

[10] Ansari, H. R., Templeton, T. J., Subudhi, A. K., Ramaprasad, A., Tang, J., Lu, F., Naeem, R., Hashish, Y., Oguike, M. C., Benavente, E. D., Clark, T. G., Sutherland, C. J., Barnwell, J. W., Culleton, R., Cao, J., & Pain, A. (2016). Genome-scale comparison of expanded gene families in *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* with *Plasmodium malariae* and with other *Plasmodium species*. *Int J Parasitol*, 46(11), 685–696.

[11] Anstey, N. M., Russell, B., Yeo, T. W., & Price, R. N. (2009). The pathophysiology of vivax malaria. *Trends in Parasitology*, 25(5), 220–227.

[12] Ariey, F., Fandeur, T., Durand, R., Randrianarivelojosia, M., Jambou, R., Legrand, E., Ekala, M. T., Bouchier, C., Cojean, S., Duchemin, J. B., Robert, V., Le Bras, J., & Mercereau-Puijalon, O. (2006). Invasion of Africa by a single *pfcrt* allele of South East Asian type. *Malaria Journal*, 5, 34.

[13] Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A. C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L., Lim, P., Leang, R., Duong, S., Sreng, S., Suon, S., Chuor, C. M., Bout, D. M., Menard, S., Rogers, W. O., Genton, B., Fandeur, T., Miotto, O., Ringwald, P., Le Bras, J., Berry, A., Barale, J. C., Fairhurst, R. M., Benoit-Vical, F., Mercereau-Puijalon, O., & Menard, D. (2014). A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature*, 505(7481), 50–5.

[14] Arisue, N., Hashimoto, T., Mitsui, H., Palacpac, N. M., Kaneko, A., Kawai, S., Hasegawa, M., Tanabe, K., & Horii, T. (2012). The *Plasmodium* apicoplast genome: conserved structure and close relationship of *P. ovale* to rodent malaria parasites. *Mol Biol Evol*, 29(9), 2095–9.

[15] Ashley, E. A., Dhorda, M., Fairhurst, R. M., Amaratunga, C., Lim, P., Suon, S., Sreng, S., Anderson, J. M., Mao, S., Sam, B., Sopha, C., Chuor, C. M., Nguon, C., Sovannaroth, S., Pukrittayakamee, S., Jittamala, P., Chotivanich, K., Chutasmit, K., Suchatsoonthorn, C., Runcharoen, R., Hien, T. T., Thuy-Nhien, N. T., Thanh, N. V., Phu, N. H., Htut, Y., Han, K. T., Aye, K. H., Mokuolu, O. A., Olaosebikan, R. R., Folaranmi, O. O., Mayxay, M., Khanthavong, M., Hongvanthong, B., Newton, P. N., Onyamboko, M. A., Fanello, C. I., Tshefu, A. K., Mishra, N., Valecha, N., Phyo, A. P., Nosten, F., Yi, P., Tripura, R., Borrmann, S., Bashraheil, M., Peshu, J., Faiz, M. A., Ghose, A., Hossain, M. A., Samad, R., Rahman, M. R., Hasan, M. M., Islam, A., Miotto, O., Amato, R., MacInnis, B., Stalker, J., Kwiatkowski, D. P.,

Bozdech, Z., Jeeyapant, A., Cheah, P. Y., Sakulthaew, T., Chalk, J., Intharabut, B., Silamut, K., Lee, S. J., Vihokhern, B., Kunasol, C., Imwong, M., Tarning, J., Taylor, W. J., Yeung, S., Woodrow, C. J., Flegg, J. A., Das, D., Smith, J., Venkatesan, M., Plowe, C. V., Stepniewska, K., Guerin, P. J., Dondorp, A. M., Day, N. P., & White, N. J. (2014). Spread of artemisinin resistance in *Plasmodium falciparum* malaria. *New England Journal of Medicine*, 371(5), 411–423.

[16] Ashley, E. A., Pyae Phyo, A., & Woodrow, C. J. (2018). Malaria. *The Lancet*, 391(10130), 1608–1621.

[17] Ashley, E. A., Stepniewska, K., Lindegardh, N., Annerberg, A., Kham, A., Brockman, A., Singhasivanon, P., White, N. J., & Nosten, F. (2007). How much fat is necessary to optimize lumefantrine oral bioavailability? *Trop Med Int Health*, 12(2), 195–200.

[18] Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., & Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15), 1968–9.

[19] Assennato, S. M., Berzuini, A., Foglieni, B., Spreafico, M., Allain, J. P., & Prati, D. (2014). *Plasmodium* genome in blood donors at risk for malaria after several years of residence in Italy. *Transfusion*, 54(10), 2419–24.

[20] Auburn, S., Bohme, U., Steinbiss, S., Trimarsanto, H., Hostetler, J., Sanders, M., Gao, Q., Nosten, F., Newbold, C. I., Berriman, M., Price, R. N., & Otto, T. D. (2016). A new *Plasmodium vivax* reference sequence with improved assembly of the subtelomeres reveals an abundance of *pir* genes. *Wellcome Open Res*, 1, 4.

[21] Auburn, S., Campino, S., Clark, T. G., Djimde, A. A., Zongo, I., Pinches, R., Manske, M., Mangano, V., Alcock, D., Anastasi, E., Maslen, G., Macinnis, B., Rockett, K., Modiano, D., Newbold, C. I., Doumbo, O. K., Ouedraogo, J. B., & Kwiatkowski, D. P. (2011). An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS One*, 6(7), e22213.

[22] Bakouh, N., Bellanca, S., Nyboer, B., Moliner Cubel, S., Karim, Z., Sanchez, C. P., Stein, W. D., Planelles, G., & Lanzer, M. (2017). Iron is a substrate of the *Plasmodium falciparum* chloroquine resistance transporter PfCRT in Xenopus oocytes. *Journal of Biological Chemistry*, 292, 16109–16121.

[23] Bannister, L., Hopkins, J., Fowler, R., Krishna, S., & Mitchell, G. (2000). A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages. *Parasitology Today*, 16(10), 427 – 433.

[24] Barber, B. E., Rajahram, G. S., Grigg, M. J., William, T., & Anstey, N. M. (2017). World malaria report: time to acknowledge *Plasmodium knowlesi* malaria. *Malaria Journal*, 16(1), 135.

[25] Barry, A. E., Leliwa-Sytek, A., Tavul, L., Imrie, H., Migot-Nabias, F., Brown, S. M., McVean, G. A. V., & Day, K. P. (2007). Population genomics of the immune evasion (*var*) genes of *Plasmodium falciparum*. *PLOS Pathogens*, 3(3), e34–.

[26] Baruch, D. I., Pasloske, B. L., Singh, H. B., Bi, X., Ma, X. C., Feldman, M., Taraschi, T. F., & Howard, R. J. (1995). Cloning the *P. falciparum* gene encoding pfemp1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1), 77–87.

[27] Bastian M., Heymann S., . J. M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

[28] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E. Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C.,

Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., & Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53 EP –.

[29] Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol*, 60(3), 291–302.

[30] Bernabeu, M., Lopez, F. J., Ferrer, M., Martin-Jaular, L., Razaname, A., Corradin, G., Maier, A. G., del Portillo, H. A., & Fernandez-Becerra, C. (2011). Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. *Cellular Microbiology*, 14(3), 386–400.

[31] Betson, M., Sousa-Figueiredo, J. C., Atuhaire, A., Arinaitwe, M., Adriko, M., Mwesigwa, G., Nabonge, J., Kabatereine, N. B., Sutherland, C. J., & Stothard, J. R. (2014). Detection of persistent *Plasmodium spp.* infections in Ugandan children after artemether-lumefantrine treatment. *Parasitology*, 141(14), 1880–90.

[32] Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., & Gething, P. W. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526, 207 EP –.

[33] Billingsley, P. & Sinden, R. (1997). Determinants of malaria-mosquito specificity. *Parasitology Today*, 13(8), 297 – 301.

[34] Billker, O., Lindo, V., Panico, M., Etienne, A. E., Paxton, T., Dell, A., Rogers, M., Sinden, R. E., & Morris, H. R. (1998). Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature*, 392, 289 EP –.

[35] Blackman, M. J. (2008). Malarial proteases and host cell egress: an 'emerging' cascade. *Cellular Microbiology*, 10(10), 1925–1934.

[36] Blasco, B., Leroy, D., & Fidock, D. A. (2017). Antimalarial drug resistance: linking *Plasmodium falciparum* parasite biology to the clinic. *Nature medicine*, 23(8), 917–928.

[37] Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., & Pirovano, W. (2011). Scaffolding preassembled contigs using SSPACE. *Bioinformatics*, 27(4), 578–9.

[38] Böhme, U., Otto, T. D., Cotton, J. A., Steinbiss, S., Sanders, M., Oyola, S. O., Nicot, A., Gandon, S., Patra, K. P., Herd, C., Bushell, E., Modrzynska, K. K., Billker, O., Vinetz, J. M., Rivero, A., Newbold, C. I., & Berriman, M. (2018). Complete avian malaria parasite genomes reveal features associated with lineage-specific evolution in birds and mammals. *Genome Research*, 28(4), 547–560.

[39] Boudin, C., Robert, V., Verhave, J. P., Carnevale, P., & Ambroise-Thomas, P. (1991). *Plasmodium falciparum* and *P. malariae* epidemiology in a West African village. *Bull World Health Organ*, 69(2), 199–205.

[40] Bousema, T., Okell, L., Felger, I., & Drakeley, C. (2014). Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nat Rev Microbiol*, 12(12), 833–40.

[41] Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T., Gentles, S., Gwilliam, R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P., Jagels, K., Jassal, B., Kyes, S., McLean, J., Moule, S., Mungall, K., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutter, S., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Whitehead, S., Woodward, J. R., Newbold, C., & Barrell, B. G. (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature*, 400, 532 EP –.

[42] Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., & Pachter, L. (2009). Fast statistical alignment. *PLoS Comput Biol*, 5(5), e1000392.

[43] Bronner, I. F., Quail, M. A., Turner, D. J., & Swerdlow, H. (2014). Improved protocols for Illumina sequencing. *Curr Protoc Hum Genet*, 80, 18 2 1–42.

[44] Brugat, T., Reid, A. J., Lin, J., Cunningham, D., Tumwine, I., Kushinga, G., McLaughlin, S., Spence, P., Böhme, U., Sanders, M., Conteh, S., Bushell, E., Metcalf, T., Billker, O., Duffy, P. E., Newbold, C., Berriman, M., & Langhorne, J. (2017). Antibody-independent mechanisms regulate the establishment of chronic *Plasmodium* infection. *Nature microbiology*, 2, 16276–16276.

[45] Bryceson, A., Fakunle, Y. M., Fleming, A. F., Crane, G., Hutt, M. S., de Cock, K. M., Greenwood, B. M., Marsden, P., & Ress, P. (1983). Malaria and splenomegaly. *Trans. R. Soc. Trop. Med. Hyg.*, 77(6), 879.

[46] Calleri, G., Balbiano, R., & Caramello, P. (2013). Are artemisinin-based combination therapies effective against *Plasmodium malariae*? *J Antimicrob Chemother*, 68(6), 1447–8.

[47] Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O., & Llinas, M. (2010). Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog*, 6(10), e1001165.

[48] Camus, D. & Hadley, T. (1985). A *Plasmodium falciparum* antigen that binds to host erythrocytes and merozoites. *Science*, 230(4725), 553–556.

[49] Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–3.

[50] Carlson, J., Helmby, H., Wahlgren, M., Carlson, J., Helmby, H., Wahlgren, M., Hill, A., Brewster, D., & Greenwood, B. M. (1990). Human cerebral malaria: association with erythrocyte rosetting and lack of anti-rosetting antibodies. *The Lancet*, 336(8729), 1457–1460.

[51] Carlton, J. M., Adams, J. H., Silva, J. C., Bidwell, S. L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S. V., Merino, E. F., Amedeo, P., Cheng, Q., Coulson, R. M., Crabb, B. S., Del Portillo, H. A., Essien, K., Feldblyum, T. V., Fernandez-Becerra, C., Gilson, P. R., Gueye, A. H., Guo, X., Kang'a, S., Kooij, T. W., Korsinczky, M., Meyer, E. V., Nene, V., Paulsen, I., White, O., Ralph, S. A., Ren, Q., Sargeant, T. J., Salzberg, S. L., Stoeckert, C. J., Sullivan, S. A., Yamamoto, M. M., Hoffman, S. L., Wortman, J. R., Gardner, M. J., Galinski, M. R., Barnwell, J. W., & Fraser-Liggett, C. M. (2008). Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*, 455(7214), 757–63.

[52] Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J., & Carucci, D. J. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, 419, 512 EP –.

[53] Carter, L. M., Kafsack, B. F., Llinás, M., Mideo, N., Pollitt, L. C., & Reece, S. E. (2013). Stress and sex in malaria parasites: Why does commitment vary? *Evolution, Medicine, and Public Health*, 2013(1), 135–147.

[54] Carter, R. (2003). Speculations on the origins of *Plasmodium vivax* malaria. *Trends in Parasitology*, 19(5), 214–219.

[55] Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the artemis comparison tool. *Bioinformatics*, 21(16), 3422–3.

[56] Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7.

[57] Cheeseman, I. H., Miller, B. A., Nair, S., Nkhoma, S., Tan, A., Tan, J. C., Saai, S. A., Phyo, A. P., Moo, C. L., Lwin, K. M., McGready, R., Ashley, E., Imwong, M., Stepniewska, K., Yi, P., Dondorp, A. M., Mayxay, M., Newton, P. N., White, N. J., Nosten, F., Ferdig, M. T., & Anderson, T. J. (2012). A major genome region underlying artemisinin resistance in malaria. *Science*, 336(6077), 79–82.

[58] Chen, Q., Barragan, A., Fernandez, V., Sundström, A., Schlichtherle, M., Sahlén, A., Carlson, J., Datta, S., & Wahlgren, M. (1998). Identification of *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP1) as the rosetting ligand of the malaria parasite *P. falciparum*. *The Journal of Experimental Medicine*, 187(1), 15–23.

[59] Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*, 10(6), 563–9.

[60] Chin, W., Contacos, P. G., Coatney, G. R., & Kimball, H. R. (1965). A naturally acquited quotidian-type malaria in man transferable to monkeys. *Science*, 149(3686), 865.

[61] Chugh, M., Sundararaman, V., Kumar, S., Reddy, V. S., Siddiqui, W. A., Stuart, K. D., & Malhotra, P. (2013). Protein complex directs hemoglobin-to-hemozoin formation in *Plasmodium falciparum*. *Proc Natl Acad Sci U S A*, 110(14), 5392–5397.

[62] Cibulskis, R. E., Alonso, P., Aponte, J., Aregawi, M., Barrette, A., Bergeron, L., Fergus, C. A., Knox, T., Lynch, M., Patouillard, E., Schwarte, S., Stewart, S., & Williams, R. (2016). Malaria: Global progress 2000 – 2015 and future challenges. *Infectious Diseases of Poverty*, 5(1), 61.

[63] Claessens, A., Hamilton, W. L., Kekre, M., Otto, T. D., Faizullabhoy, A., Rayner, J. C., & Kwiatkowski, D. (2014). Generation of antigenic diversity in *Plasmodium falciparum* by structured rearrangement of Var genes during mitosis. *PLoS Genet*, 10(12), e1004812.

[64] Claudio, J. O., Liew, C. C., Ma, J., Heng, H. H., Stewart, A. K., & Hawley, R. G. (1999). Cloning and expression analysis of a novel WD repeat gene, WDR3, mapping to 1p12-p13. *Genomics*, 59(1), 85–9.

[65] Clausen, T. M., Christoffersen, S., Dahlbäck, M., Langkilde, A. E., Jensen, K. E., Resende, M., Agerbæk, M. Ø., Andersen, D., Berisha, B., Ditlev, S. B., Pinto, V. V., Nielsen, M. A., Theander, T. G., Larsen, S., & Salanti, A. (2012). Structural and functional insight into how the *Plasmodium falciparum* VAR2CSA protein mediates binding to chondroitin sulfate A in placental malaria. *The Journal of Biological Chemistry*, 287(28), 23332–23345.

[66] Cockburn, I. A., Mackinnon, M. J., O'Donnell, A., Allen, S. J., Moulds, J. M., Baisor, M., Bockarie, M., Reeder, J. C., & Rowe, J. A. (2004). A human complement receptor 1 polymorphism that reduces *Plasmodium falciparum* rosetting confers protection against severe malaria. *Proc Natl Acad Sci U S A*, 101(1), 272–277.

[67] Collins, W. E. & Jeffery, G. M. (2005). *Plasmodium ovale*: parasite and disease. *Clin Microbiol Rev*, 18(3), 570–81.

[68] Collins, W. E. & Jeffery, G. M. (2007). *Plasmodium malariae*: parasite and disease. *Clin Microbiol Rev*, 20(4), 579–92.

[69] Cowman, A. F. & Crabb, B. S. (2002). The *Plasmodium falciparum* genome–a blueprint for erythrocyte invasion. *Science*, 298(5591), 126–8.

[70] Cowman, A. F., Galatis, D., & Thompson, J. K. (1994). Selection for mefloquine resistance in *Plasmodium falciparum* is linked to amplification of the *pfmdr1* gene and cross-resistance to halofantrine and quinine. *Proc Natl Acad Sci U S A*, 91(3), 1143–7.

[71] Cowman, A. F., Tonkin, C. J., Tham, W.-H., & Duraisingh, M. T. (2017). The molecular basis of erythrocyte invasion by malaria parasites. *Cell Host & Microbe*, 22(2), 232–245.

[72] Cox-Singh, J., Davis, T. M. E., Lee, K.-S., Shamsul, S. S. G., Matusop, A., Ratnam, S., Rahman, H. A., Conway, D. J., & Singh, B. (2008). *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. *Clinical Infectious Diseases*, 46(2), 165–171.

[73] Craig, A. G., Grau, G. E., Janse, C., Kazura, J. W., Milner, D., Barnwell, J. W., Turner, G., Langhorne, J., & on behalf of the participants of the Hinxton Retreat meeting on "Animal

Models for Research on Severe Malaria" (2012). The role of animal models for research on severe malaria. *PLoS Pathogens*, 8(2), e1002401.

[74] Crosnier, C., Bustamante, L. Y., Bartholdson, S. J., Bei, A. K., Theron, M., Uchikawa, M., Mboup, S., Ndir, O., Kwiatkowski, D. P., Duraisingh, M. T., Rayner, J. C., & Wright, G. J. (2011). Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*, 480(7378), 534–7.

[75] Culleton, R. & Carter, R. (2012). African *Plasmodium vivax*: Distribution and origins. *International Journal for Parasitology*, 42(12), 1091 – 1097. Singapore Malaria Network Meeting (SingMalNet) 2012.

[76] Cunningham, D., Lawton, J., Jarra, W., Preiser, P., & Langhorne, J. (2010). The *pir* multigene family of *Plasmodium*: antigenic variation and beyond. *Mol Biochem Parasitol*, 170(2), 65–73.

[77] Daniels, R. F., Schaffner, S. F., Wenger, E. A., Proctor, J. L., Chang, H. H., Wong, W., Baro, N., Ndiaye, D., Fall, F. B., Ndiop, M., Ba, M., Milner, D. A., J., Taylor, T. E., Neafsey, D. E., Volkman, S. K., Eckhoff, P. A., Hartl, D. L., & Wirth, D. F. (2015). Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proc Natl Acad Sci U S A*, 112(22), 7067–72.

[78] de Koning-Ward, T. F., Dixon, M. W. A., Tilley, L., & Gilson, P. R. (2016). *Plasmodium* species: master renovators of their host cells. *Nature Reviews Microbiology*, 14, 494 EP –.

[79] Despommier, D., Gwadz, R. W., Hotez, P. J., & Knirsch, C. A. (2000). *Parasitic Diseases*. New York: Apple Trees Productions.

[80] Dixon, M. W. A., Thompson, J., Gardiner, D. L., & Trenholme, K. R. (2008). Sex in *Plasmodium*: a sign of commitment. *Trends in Parasitology*, 24(4), 168–175.

[81] Djimdé, A., Doumbo, O. K., Cortese, J. F., Kayentao, K., Doumbo, S., Diourté, Y., Coulibaly, D., Dicko, A., Su, X.-z., Nomura, T., Fidock, D. A., Wellems, T. E., & Plowe, C. V. (2001). A molecular marker for chloroquine-resistant falciparum malaria. *New England Journal of Medicine*, 344(4), 257–263.

[82] Djimde, A., Doumbo, O. K., Steketee, R. W., & Plowe, C. V. (2001). Application of a molecular marker for surveillance of chloroquine-resistant falciparum malaria. *Lancet*, 358(9285), 890–891.

[83] Doderer-Lang, C., Atchade, P. S., Meckert, L., Haar, E., Perrotey, S., Filisetti, D., Aboubacar, A., Pfaff, A. W., Brunet, J., Chabi, N. W., Akpovi, C. D., Anani, L., Bigot, A., Sanni, A., &

Candolfi, E. (2014). The ears of the African elephant: unexpected high seroprevalence of *Plasmodium ovale* and *Plasmodium malariae* in healthy populations in Western Africa. *Malar J*, 13, 240.

[84] Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J., Ringwald, P., Silamut, K., Imwong, M., Chotivanich, K., Lim, P., Herdman, T., An, S. S., Yeung, S., Singhasivanon, P., Day, N. P. J., Lindegardh, N., Socheat, D., & White, N. J. (2009). Artemisinin resistance in *Plasmodium falciparum* malaria. *New England Journal of Medicine*, 361(5), 455–467.

[85] Doolittle, R. F. (2002). The grand assault. *Nature*, 419, 493 EP –.

[86] Douglas, N. M., Anstey, N. M., Angus, B. J., Nosten, F., & Price, R. N. (2010). Artemisinin combination therapy for vivax malaria. *Lancet Infect Dis*, 10(6), 405–16.

[87] Douglas, N. M., Lampah, D. A., Kenangalem, E., Simpson, J. A., Poespoprodjo, J. R., Sugiarto, P., Anstey, N. M., & Price, R. N. (2013). Major burden of severe anemia from non-falciparum malaria species in Southern Papua: a hospital-based surveillance study. *PLoS Med*, 10(12), e1001575.

[88] Droege, M. & Hill, B. (2008). The genome sequencer flx™ system—longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*, 136(1), 3 – 10.

[89] Duraisingh, M. T., Roper, C., Walliker, D., & Warhurst, D. C. (2002). Increased sensitivity to the antimalarials mefloquine and artemisinin is conferred by mutations in the *pfmdr1* gene of *Plasmodium falciparum*. *Molecular Microbiology*, 36(4), 955–961.

[90] Durrand, V., Berry, A., Sem, R., Glaziou, P., Beaudou, J., & Fandeur, T. (2004). Variations in the sequence and expression of the *Plasmodium falciparum* chloroquine resistance transporter (*Pfcrt*) and their relationship to chloroquine resistance *in vitro*. *Molecular and Biochemical Parasitology*, 136(2), 273–285.

[91] Dyer, M. & Day, K. P. (2003). Regulation of the rate of asexual growth and commitment to sexual development by diffusible factors from *in vitro* cultures of *Plasmodium falciparum*. *Am. J. Trop. Med. Hyg.*, 68(4), 403 – 409.

[92] Eastman, R. T., Dharia, N. V., Winzeler, E. A., & Fidock, D. A. (2011). Piperaquine resistance is associated with a copy number variation on chromosome 5 in drug-pressured *Plasmodium falciparum* parasites. *Antimicrobial Agents and Chemotherapy*, 55(8), 3908–3916.

[93] Ecker, A., Lehane, A. M., Clain, J., & Fidock, D. A. (2012). PfCRT and its role in antimalarial drug resistance. *Trends in Parasitology*, 28(11), 504–514.

[94] Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5), 1792–7.

[95] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., & Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), 133–138.

[96] Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7), 1575–84.

[97] Escalante, A. A. & Ayala, F. J. (1994). Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. *Proc Natl Acad Sci U S A*, 91(24), 11373–11377.

[98] Escalante, A. A., Lal, A. A., & Ayala, F. J. (1998). Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. *Genetics*, 149(1), 189–202.

[99] Ezzet, F., van Vugt, M., Nosten, F., Looareesuwan, S., & White, N. J. (2000). Pharmacokinetics and pharmacodynamics of lumefantrine (benflumetol) in acute falciparum malaria. *Antimicrob Agents Chemother*, 44(3), 697–704.

[100] Fan, X., Chaisson, M., Nakhleh, L., & Chen, K. (2017). HySA: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. *Genome Research*, 27(5), 793–800.

[101] Feachem, R. & Sabot, O. (2008). A new global malaria eradication strategy. *The Lancet*, 371(9624), 1633–1635.

[102] Ferreira, M. U., da Silva Nunes, M., & Wunderlich, G. (2004). Antigenic diversity and immune evasion by malaria parasites. *Clinical and Diagnostic Laboratory Immunology*, 11(6), 987–995.

[103] Fidock, D. A., Nomura, T., Talley, A. K., Cooper, R. A., Dzekunov, S. M., Ferdig, M. T., Ursos, L. M., Sidhu, A. B., Naude, B., Deitsch, K. W., Su, X. Z., Wootton, J. C., Roepe, P. D., & Wellems, T. E. (2000). Mutations in the *P. falciparum* digestive vacuole transmembrane protein PfCRT and evidence for their role in chloroquine resistance. *Mol Cell*, 6(4), 861–71.

[104] Flegg, J. A., Guerin, P. J., White, N. J., & Stepniewska, K. (2011). Standardizing the measurement of parasite clearance in falciparum malaria: the parasite clearance estimator. *Malaria Journal*, 10, 339–339.

[105] Foote, S. J., Kyle, D. E., Martin, R. K., Oduola, A. M. J., Forsyth, K., Kemp, D. J., & Cowman, A. F. (1990). Several alleles of the multidrug-resistance gene are closely linked to chloroquine resistance in *Plasmodium falciparum*. *Nature*, 345, 255 EP –.

[106] Fougere, A., Jackson, A. P., Bechtsi, D. P., Braks, J. A., Annoura, T., Fonager, J., Spaccapelo, R., Ramesar, J., Chevalley-Maurel, S., Klop, O., van der Laan, A. M., Tanke, H. J., Kocken, C. H., Pasini, E. M., Khan, S. M., Bohme, U., van Ooij, C., Otto, T. D., Janse, C. J., & Franke-Fayard, B. (2016). Variant exported blood-stage proteins encoded by *Plasmodium* multigene families are expressed in liver stages where they are exported into the parasitophorous vacuole. *PLoS Pathog*, 12(11), e1005917.

[107] Franken, G., Muller-Stover, I., Holtfreter, M. C., Walter, S., Mehlhorn, H., Labisch, A., Haussinger, D., & Richter, J. (2012). Why do *Plasmodium malariae* infections sometimes occur in spite of previous antimalarial medication? *Parasitol Res*, 111(2), 943–6.

[108] Frech, C. & Chen, N. (2013). Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. *BMC Genomics*, 14, 427.

[109] Fruchterman, T. M. J. & Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129–1164.

[110] Fuehrer, H.-P., Habler, V. E., Fally, M. A., Harl, J., Starzengruber, P., Swoboda, P., Bloeschl, I., Khan, W. A., & Noedl, H. (2012). *Plasmodium ovale* in Bangladesh: Genetic diversity and the first known evidence of the sympatric distribution of *Plasmodium ovale curtisi* and *Plasmodium ovale wallikeri* in southern Asia. *International Journal for Parasitology*, 42(7), 693 – 699.

[111] Galaway, F., Drought, L. G., Fala, M., Cross, N., Kemp, A. C., Rayner, J. C., & Wright, G. J. (2017). P113 is a merozoite surface protein that binds the N terminus of *Plasmodium falciparum* RH5. *Nature Communications*, 8, 14333.

[112] Galen, S. C., Borner, J., Martinsen, E. S., Schaer, J., Austin, C. C., West, C. J., & Perkins, S. L. (2018). The polyphyly of *Plasmodium*: comprehensive phylogenetic analyses of the malaria parasites (order Haemosporida) reveal widespread taxonomic conflict. *Royal Society Open Science*, 5(5).

[113] Galinski, M. R., Medina, C. C., Ingravallo, P., & Barnwell, J. W. (1992). A reticulocyte-binding protein complex of *Plasmodium vivax* merozoites. *Cell*, 69(7), 1213–26.

[114] Garcia, C. R. S., Markus, R. P., & Madeira, L. (2001). Tertian and quartan fevers: Temporal regulation in malarial infection. *Journal of Biological Rhythms*, 16(5), 436–443.

[115] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M. S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., & Barrell, B. (2002a). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906), 498–511.

[116] Gardner, M. J., Shallom, S. J., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L. M., Smith, H. O., Adams, M. D., Venter, J. C., Carucci, D. J., Hoffman, S. L., & Fraser, C. M. (2002b). Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature*, 419, 531 EP –.

[117] Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., Sutton, G. G., Clayton, R., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C., & Hoffman, S. L. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science*, 282(5391), 1126–1132.

[118] Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L., Tatem, A. J., & Hay, S. I. (2011). A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal*, 10(1), 378.

[119] Gilles, H. M. & Hendrickse, R. G. (1963). Nephrosis in Nigerian children. role of *Plasmodium malariae*, and effect of antimalarial treatment. *Br Med J*, 2(5348), 27–31.

[120] Gilson, P. R. & Crabb, B. S. (2009). Morphology and kinetics of the three distinct phases of red blood cell invasion by *Plasmodium falciparum* merozoites. *International Journal for Parasitology*, 39(1), 91 – 96.

[121] Goel, S., Palmkvist, M., Moll, K., Joannin, N., Lara, P., R Akhouri, R., Moradi, N., Öjemalm, K., Westman, M., Angeletti, D., Kjellin, H., Lehtiö, J., Blixt, O., Ideström, L., Gahmberg, C. G., Storry, J. R., Hult, A. K., Olsson, M. L., von Heijne, G., Nilsson, I., & Wahlgren, M. (2015). RIFINs are adhesins implicated in severe *Plasmodium falciparum* malaria. *Nature Medicine*, 21, 314 EP –.

[122] Gomes, P. S., Bhardwaj, J., Rivera-Correa, J., Freire-De-Lima, C. G., & Morrot, A. (2016). Immune escape strategies of malaria parasites. *Frontiers in Microbiology*, 7, 1617.

[123] Greenwood, B. (2017). Elimination of malaria: halfway there. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 111(1), 1–2.

[124] Grigg, M. J., William, T., Menon, J., Dhanaraj, P., Barber, B. E., Wilkes, C. S., von Seidlein, L., Rajahram, G. S., Pasay, C., McCarthy, J. S., Price, R. N., Anstey, N. M., & Yeo, T. W. (2016). Artesunate-mefloquine versus chloroquine for treatment of uncomplicated *Plasmodium knowlesi* malaria in malaysia (ACT KNOW): an open-label, randomised controlled trial. *Lancet Infect Dis*, 16(2), 180–8.

[125] Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet*, 43(10), 1031–4.

[126] Gruning, B. A., Rasche, E., Rebolledo-Jaramillo, B., Eberhard, C., Houwaart, T., Chilton, J., Coraor, N., Backofen, R., Taylor, J., & Nekrutenko, A. (2017). Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers. *Plos Computational Biology*, 13(5).

[127] Gruszczyk, J., Lim, N. T., Arnott, A., He, W. Q., Nguitragool, W., Roobsoong, W., Mok, Y. F., Murphy, J. M., Smith, K. R., Lee, S., Bahlo, M., Mueller, I., Barry, A. E., & Tham, W. H. (2016). Structurally conserved erythrocyte-binding domain in *Plasmodium* provides a versatile scaffold for alternate receptor engagement. *Proc Natl Acad Sci U S A*, 113(2), E191–200.

[128] Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, 59(3), 307–21.

[129] Guo, S., Kyaw, M. P., He, L., Min, M., Ning, X., Zhang, W., Wang, B., & Cui, L. (2017). Quality testing of artemisinin-based antimalarial drugs in Myanmar. *The American Journal of Tropical Medicine and Hygiene*, 97(4), 1198–1203.

[130] Haldar, K. & Mohandas, N. (2009). Malaria, erythrocytic infection, and anemia. *Hematology*, (pp. 87–93).

[131] Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M. A., Ormond, D., Doggett, J., Trueman, H. E., Mendoza, J., Bidwell, S. L., Rajandream, M.-A., Carucci, D. J., Yates, J. R., Kafatos, F. C., Janse, C. J., Barrell, B., Turner, C. M. R., Waters, A. P., & Sinden, R. E. (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science*, 307(5706), 82–86.

[132] Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S., Barron, A., Brooks, K., Buckee, C. O., Burrows, C., Cherevach, I., Chillingworth, C., Chillingworth, T., Christodoulou, Z., Clark, L., Clark, R., Corton, C., Cronin, A., Davies, R., Davis, P., Dear, P., Dearden, F., Doggett, J., Feltwell, T., Goble, A., Goodhead, I., Gwilliam, R., Hamlin, N., Hance, Z., Harper, D., Hauser, H., Hornsby, T., Holroyd, S., Horrocks, P., Humphray, S., Jagels, K., James, K. D., Johnson, D., Kerhornou, A., Knights, A., Konfortov, B., Kyes, S., Larke, N., Lawson, D., Lennard, N., Line, A., Maddison, M., McLean, J., Mooney, P., Moule, S., Murphy, L., Oliver, K., Ormond, D., Price, C., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Rutter, S., Rutherford, K. M., Sanders, M., Simmonds, M., Seeger, K., Sharp, S., Smith, R., Squares, R., Squares, S., Stevens, K., Taylor, K., Tivey, A., Unwin, L., Whitehead, S., Woodward, J., Sulston, J. E., Craig, A., Newbold, C., & Barrell, B. G. (2002). Sequence of *Plasmodium falciparum* chromosomes 1, 3–9 and 13. *Nature*, 419, 527 EP –.

[133] Hartl, D. L., Volkman, S. K., Nielsen, K. M., Barry, A. E., Day, K. P., Wirth, D. F., & Winzeler, E. A. (2002). The paradoxical population genetics of *Plasmodium falciparum*. *Trends in Parasitology*, 18(6), 266 – 272.

[134] Hastings, I. (2011). How artemisinin-containing combination therapies slow the spread of antimalarial drug resistance. *Trends in Parasitology*, 27(2), 67 – 72.

[135] Hay, S. I., Guerra, C. A., Tatem, A. J., Noor, A. M., & Snow, R. W. (2004). The global distribution and population at risk of malaria: past, present, and future. *The Lancet Infectious Diseases*, 4(6), 327–336.

[136] Hay, S. I., Sinka, M. E., Okara, R. M., Kabaria, C. W., Mbithi, P. M., Tago, C. C., Benz, D., Gething, P. W., Howes, R. E., Patil, A. P., Temperley, W. H., Bangs, M. J., Chareonviriyaphap, T., Elyazar, I. R. F., Harbach, R. E., Hemingway, J., Manguin, S., Mbogo, C. M., Rubio-Palis, Y., & Godfray, H. C. J. (2010). Developing global maps of the dominant *Anopheles* vectors of human malaria. *PLOS Medicine*, 7(2), 1–6.

[137] Haynes, J. D., Dalton, J. P., Klotz, F. W., McGinniss, M. H., Hadley, T. J., Hudson, D. E., & Miller, L. H. (1988). Receptor-like specificity of a *Plasmodium knowlesi* malarial protein that binds to Duffy antigen ligands on erythrocytes. *The Journal of Experimental Medicine*, 167(6), 1873–1881.

[138] Hemingway, J., Ranson, H., Magill, A., Kolaczinski, J., Fornadel, C., Gimnig, J., Coetzee, M., Simard, F., Roch, D. K., Hinzoumbe, C. K., Pickett, J., Schellenberg, D., Gething, P., Hoppé, M., & Hamon, N. (2016). Averting a malaria disaster: will insecticide resistance derail malaria control? *The Lancet*, 387(10029), 1785 – 1788.

[139] Hiller, N. L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C., & Haldar, K. (2004). A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science*, 306(5703), 1934–7.

[140] Ho, M., Davis, T. M., Silamut, K., Bunnag, D., & White, N. J. (1991). Rosette formation of *Plasmodium falciparum*-infected erythrocytes from patients with acute malaria. *Infection and Immunity*, 59(6), 2135–2139.

[141] Hoenen, T., Groseth, A., Rosenke, K., Fischer, R. J., Hoenen, A., Judson, S. D., Martellaro, C., Falzarano, D., Marzi, A., Squires, R. B., Wollenberg, K. R., de Wit, E., Prescott, J., Safronetz, D., van Doremalen, N., Bushmaker, T., Feldmann, F., McNally, K., Bolay, F. K., Fields, B., Sealy, T., Rayfield, M., Nichol, S. T., Zoon, K. C., Massaquoi, M., Munster, V. J., & Feldmann, H. (2016). Nanopore sequencing as a rapidly deployable ebola outbreak tool. *Emerging Infectious Diseases*, 22(2), 331–334.

[142] Hoffman, S. L., Subramanian, G. M., Collins, F. H., & Venter, J. C. (2002). *Plasmodium*, human and *Anopheles* genomics and malaria. *Nature*, 415, 702 EP –.

[143] Holloway, A. K., Lawniczak, M. K., Mezey, J. G., Begun, D. J., & Jones, C. D. (2007). Adaptive gene expression divergence inferred from population genomics. *PLoS Genet*, 3(10), 2007–13.

[144] Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. C., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., Atkinson, P. W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chatuverdi, K., Christophides, G. K., Chrystal, M. A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C. A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M. E., Hladun, S. L., Hogan, J. R., Hong, Y. S., Hoover, J., Jaillon, O., Ke, Z.,

Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.-J., Lobo, N. F., Lopez, J. R., Malek, J. A., McIntosh, T. C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S. D., O'Brochta, D. A., Pfannkoch, C., Qi, R., Regier, M. A., Remington, K., Shao, H., Sharakhova, M. V., Sitter, C. D., Shetty, J., Smith, T. J., Strong, R., Sun, J., Thomasova, D., Ton, L. Q., Topalis, P., Tu, Z., Unger, M. F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K. J., Wortman, J. R., Wu, M., Yao, A., Zdobnov, E. M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S. C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C. W., Louis, C., Kalush, F., Mural, R. J., Myers, E. W., Adams, M. D., Smith, H. O., Broder, S., Gardner, M. J., Fraser, C. M., Birney, E., Bork, P., Brey, P. T., Venter, J. C., Weissenbach, J., Kafatos, F. C., Collins, F. H., & Hoffman, S. L. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591), 129–149.

[145] Hong, E. P. & Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, 10(2), 117–122.

[146] Huang, X. (1996). An improved sequence assembly program. *Genomics*, 33(1), 21 – 31.

[147] Hupalo, D. N., Luo, Z., Melnikov, A., Sutton, P. L., Rogov, P., Escalante, A., Vallejo, A. F., Herrera, S., Arévalo-Herrera, M., Fan, Q., Wang, Y., Cui, L., Lucas, C. M., Durand, S., Sanchez, J. F., Baldeviano, G. C., Lescano, A. G., Laman, M., Barnadas, C., Barry, A., Mueller, I., Kazura, J. W., Eapen, A., Kanagaraj, D., Valecha, N., Ferreira, M. U., Roobsoong, W., Nguitragool, W., Sattabonkot, J., Gamboa, D., Kosek, M., Vinetz, J. M., González-Cerón, L., Birren, B. W., Neafsey, D. E., & Carlton, J. M. (2016). Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. *Nature Genetics*, 48, 953 EP –.

[148] Hyman, R. W., Fung, E., Conway, A., Kurdi, O., Mao, J., Miranda, M., Nakao, B., Rowley, D., Tamaki, T., Wang, F., & Davis, R. W. (2002). Sequence of *Plasmodium falciparum* chromosome 12. *Nature*, 419, 534 EP –.

[149] Imwong, M., Hien, T. T., Thuy-Nhien, N. T., Dondorp, A. M., & White, N. J. (2017a). Spread of a single multidrug resistant malaria parasite lineage (PfPailin) to Vietnam. *Lancet Infect Dis*, 17(10), 1022–1023.

[150] Imwong, M., Suwannasin, K., Kunasol, C., Sutawong, K., Mayxay, M., Rekol, H., Smithuis, F. M., Hlaing, T. M., Tun, K. M., van der Pluijm, R. W., Tripura, R., Miotto, O., Menard, D., Dhorda, M., Day, N. P. J., White, N. J., & Dondorp, A. M. (2017b). The spread of artemisinin-resistant *Plasmodium falciparum* in the Greater Mekong subregion: a molecular epidemiology observational study. *Lancet Infect Dis*, 17(5), 491–497.

[151] Imwong, M., Tanomsing, N., Pukrittayakamee, S., Day, N. P., White, N. J., & Snounou, G. (2009). Spurious amplification of a *Plasmodium vivax* small-subunit RNA gene by use of primers currently used to detect *P. knowlesi*. *J Clin Microbiol*, 47(12), 4173–5.

[152] Innan, H. (2006). Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics*, 173(3), 1725–33.

[153] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860 EP –.

[154] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8), e124–.

[155] Iyer, J., Gruner, A. C., Renia, L., Snounou, G., & Preiser, P. R. (2007). Invasion of host cells by malaria parasites: a tale of two protein families. *Mol Microbiol*, 65(2), 231–49.

[156] Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., & Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36, 338 EP –.

[157] Janssen, C. S., Phillips, R. S., Turner, C. M. R., & Barrett, M. P. (2004). *Plasmodium* interspersed repeats: the major multigene superfamily of malaria parasites. *Nucleic Acids Research*, 32(19), 5712–5720.

[158] Jianbing, M., T., F. M., Xiaorong, F., A., J. D., Junhui, D., Tetsuya, F., G., S., L., A., A., C. R., C., W. J., Momiao, X., & Xin-zhuan, S. (2003). Multiple transporters associated with malaria parasite responses to chloroquine and quinine. *Molecular Microbiology*, 49(4), 977–989.

[159] Josling, G. A. & Llinás, M. (2015). Sexual development in *Plasmodium* parasites: knowing when it's time to commit. *Nature Reviews Microbiology*, 13, 573 EP –.

[160] Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., & Tse, D. N. (2017). HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research*, 27(5), 747–756.

[161] Kamau, E., Campino, S., Amenga-Etego, L., Drury, E., Ishengoma, D., Johnson, K., Mumba, D., Kekre, M., Yavo, W., Mead, D., Bouyou-Akotet, M., Apinjoh, T., Golassa, L., Randrianarivelojosia, M., Andagalu, B., Maiga-Ascofare, O., Amambua-Ngwa, A., Tindana, P., Ghansah, A., MacInnis, B., Kwiatkowski, D., & Djimde, A. A. (2015). K13-propeller polymorphisms in *Plasmodium falciparum* parasites from sub-Saharan Africa. *J Infect Dis*, 211(8), 1352–5.

[162] Kamini, M., Aafje, R., Marian, W., Andrea, B., Brian, G., & H., W. W. (2009). From malaria control to eradication: The WHO perspective. *Tropical Medicine & International Health*, 14(7), 802–809.

[163] Kaneko, A., Taleo, G., Kalkoa, M., Yaviong, J., Reeve, P. A., Ganczakowski, M., Shirakawa, C., Palmer, K., Kobayakawa, T., & Bjorkman, A. (1998). Malaria epidemiology, glucose 6-phosphate dehydrogenase deficiency and human settlement in the Vanuatu Archipelago. *Acta Trop*, 70(3), 285–302.

[164] Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15, 379 EP –.

[165] Karunaweera, N. D., Galappaththy, G. N., & Wirth, D. F. (2014). On the road to eliminate malaria in Sri Lanka: lessons from history, challenges, gaps in knowledge and research needs. *Malaria Journal*, 13(1), 59.

[166] Keeling, P. J. & Rayner, J. C. (2015). The origins of malaria: there are more things in heaven and earth. *Parasitology*, 142 Suppl 1, S16–25.

[167] Killeen, G. F. & Ranson, H. (2018). Insecticide-resistant malaria vectors must be tackled. *The Lancet*, 391(10130), 1551–1552.

[168] Klonis, N., Crespo-Ortiz, M. P., Bottova, I., Abu-Bakar, N., Kenny, S., Rosenthal, P. J., & Tilley, L. (2011). Artemisinin activity against *Plasmodium falciparum* requires hemoglobin uptake and digestion. *Proc Natl Acad Sci U S A*, 108(28), 11405–10.

[169] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & development team, J. (2016). Jupyter notebooks: a publishing format for reproducible computational workflows. In F. Loizides & B. Scmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).: IOS Press.

[170] Knope, K., Liu, C., Begg, K., Johansen, C., Whelan, P., & Kurucz, N. (2013). Communicable diseases network australia national arbovirus and malaria advisory. *Commun Dis Intell Q Re*, 32(1), 31–47.

[171] Kooij, T. W. A., Janse, C. J., & Waters, A. P. (2006). *Plasmodium* post-genomics: better the bug you know? *Nature Reviews Microbiology*, 4, 344 EP –.

[172] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), 722–736.

[173] Kraemer, S. M., Kyes, S. A., Aggarwal, G., Springer, A. L., Nelson, S. O., Christodoulou, Z., Smith, L. M., Wang, W., Levin, E., Newbold, C. I., Myler, P. J., & Smith, J. D. (2007). Patterns of gene recombination shape *var* gene repertoires in *Plasmodium falciparum*: comparisons of geographically diverse isolates. *BMC Genomics*, 8, 45–45.

[174] Kreitman, M. & Hudson, R. R. (1991). Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*, 127(3), 565–82.

[175] Krotoski, W. A. (1985). Discovery of the hypnozoite and a new theory of malarial relapse. *Trans R Soc Trop Med Hyg*, 79(1), 1 – 11.

[176] Lalloo, D. G., Shingadia, D., Bell, D. J., Beeching, N. J., Whitty, C. J., Chiodini, P. L., & Travellers, P. H. E. (2016). UK malaria treatment guidelines 2016. *J Infect*, 72(6), 635–49.

[177] Lalremruata, A., Magris, M., Vivas-Martinez, S., Koehler, M., Esen, M., Kempaiah, P., Jeyaraj, S., Perkins, D. J., Mordmuller, B., & Metzger, W. G. (2015). Natural infection of *Plasmodium brasilianum* in humans: Man and monkey share quartan malaria parasites in the Venezuelan Amazon. *EBioMedicine*, 2(9), 1186–92.

[178] Langford, S., Douglas, N. M., Lampah, D. A., Simpson, J. A., Kenangalem, E., Sugiarto, P., Anstey, N. M., Poespoprodjo, J. R., & Price, R. N. (2015). *Plasmodium malariae* infection associated with a high burden of anemia: A hospital-based surveillance study. *PLoS Negl Trop Dis*, 9(12), e0004195.

[179] Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4), 357–9.

[180] Lapp, S. A., Geraldo, J. A., Chien, J.-T., Ay, F., Pakala, S. B., Batugedara, G., Humphrey, J., the MaHPIC consortium, DeBarry, J. D., Le Roch, K. G., Galinski, M. R., & Kissinger, J. C. (2018). Pacbio assembly of a *Plasmodium knowlesi* genome sequence with Hi-C correction and manual annotation of the SICAvar gene family. *Parasitology*, 145(1), 71–84.

[181] Larremore, D. B., Sundararaman, S. A., Liu, W., Proto, W. R., Clauset, A., Loy, D. E., Speede, S., Plenderleith, L. J., Sharp, P. M., Hahn, B. H., Rayner, J. C., & Buckee, C. O. (2015). Ape parasite origins of human malaria virulence genes. *Nat Commun*, 6, 8368.

[182] Lartillot, N., Lepage, T., & Blanquart, S. (2009). Phylobayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17), 2286–8.

[183] Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., Craig, J. M., Langford, K. W., Samson, J. M., Daza, R., Doering, K., Shendure, J., & Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature biotechnology*, 32(8), 829–833.

[184] Laufer, M. K., Thesing, P. C., Eddington, N. D., Masonga, R., Dzinjalamala, F. K., Takala, S. L., Taylor, T. E., & Plowe, C. V. (2006). Return of chloroquine antimalarial efficacy in Malawi. *New England Journal of Medicine*, 355(19), 1959–1966. PMID: 17093247.

[185] Leang, R., Barrette, A., Bouth, D. M., Menard, D., Abdur, R., Duong, S., & Ringwald, P. (2013a). Efficacy of dihydroartemisinin-piperaquine for treatment of uncomplicated *Plasmodium falciparum* and *Plasmodium vivax* in Cambodia, 2008 to 2010. *Antimicrob Agents Chemother*, 57(2), 818–26.

[186] Leang, R., Ros, S., Duong, S., Navaratnam, V., Lim, P., Ariey, F., Kiechel, J.-R., Ménard, D., & Taylor, W. R. (2013b). Therapeutic efficacy of fixed dose artesunate-mefloquine for the treatment of acute, uncomplicated *Plasmodium falciparum* malaria in Kampong Speu, Cambodia. *Malaria Journal*, 12(1), 343.

[187] Leang, R., Taylor, W. R., Bouth, D. M., Song, L., Tarning, J., Char, M. C., Kim, S., Witkowski, B., Duru, V., Domergue, A., Khim, N., Ringwald, P., & Menard, D. (2015). Evidence of *Plasmodium falciparum* malaria multidrug resistance to artemisinin and piperaquine in Western Cambodia: Dihydroartemisinin-piperaquine open-label multicenter clinical assessment. *Antimicrob Agents Chemother*, 59(8), 4719–26.

[188] Leichty, A. R. & Brisson, D. (2014). Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics*, 198(2), 473–481.

[189] Levine, N. D. (2017). *The Protozoan Phylum Apicomplexa*, volume 1. Boca Raton: Taylor & Francis Group.

[190] Lewis, I. A., Wacker, M., Olszewski, K. L., Cobbold, S. A., Baska, K. S., Tan, A., Ferdig, M. T., & Llinas, M. (2014). Metabolic QTL analysis links chloroquine resistance in *Plasmodium falciparum* to impaired hemoglobin catabolism. *PLoS Genet*, 10(1), e1004085.

[191] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.

[192] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–9.

[193] Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11), 1851–1858.

[194] Li, L., Stoeckert, C. J., J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9), 2178–89.

[195] Li, W., Boswell, R., & Wood, W. B. (2000). mag-1, a homolog of *Drosophila* mago nashi, regulates hermaphrodite germ-line sex determination in *Caenorhabditis elegans*. *Dev Biol*, 218(2), 172–82.

[196] Lim, L. & McFadden, G. I. (2010). The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1541), 749–763.

[197] Lim, P., Chy, S., Ariey, F., Incardona, S., Chim, P., Sem, R., Denis, M. B., Hewitt, S., Hoyer, S., Socheat, D., Merecreau-Puijalon, O., & Fandeur, T. (2003). *pfcrt* polymorphism and chloroquine resistance in *Plasmodium falciparum* strains isolated in Cambodia. *Antimicrobial Agents and Chemotherapy*, 47(1), 87–94.

[198] Lim, P., Dek, D., Try, V., Eastman, R. T., Chy, S., Sreng, S., Suon, S., Mao, S., Sopha, C., Sam, B., Ashley, E. A., Miotto, O., Dondorp, A. M., White, N. J., Su, X.-z., Char, M. C., Anderson, J. M., Amaratunga, C., Menard, D., & Fairhurst, R. M. (2013). *Ex Vivo* susceptibility of *Plasmodium falciparum* to antimalarial drugs in Western, Northern, and Eastern Cambodia, 2011-2012: Association with molecular markers. *Antimicrobial Agents and Chemotherapy*, 57(11), 5277–5283.

[199] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8(10), 833–U94.

[200] Liu, W., Li, Y., Learn, G. H., Rudicell, R. S., Robertson, J. D., Keele, B. F., Ndjango, J.-B. N., Sanz, C. M., Morgan, D. B., Locatelli, S., Gonder, M. K., Kranzusch, P. J., Walsh, P. D., Delaporte, E., Mpoudi-Ngole, E., Georgiev, A. V., Muller, M. N., Shaw, G. M., Peeters, M., Sharp, P. M., Rayner, J. C., & Hahn, B. H. (2010). Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. *Nature*, 467, 420 EP –.

[201] Lobo, C. A., Fujioka, H., Aikawa, M., & Kumar, N. (1999). Disruption of the Pfg27 locus by homologous recombination leads to loss of the sexual phenotype in *P. falciparum*. *Mol Cell*, 3(6), 793–8.

[202] Lopaticki, S., Maier, A. G., Thompson, J., Wilson, D. W., Tham, W.-H., Triglia, T., Gout, A., Speed, T. P., Beeson, J. G., Healer, J., & Cowman, A. F. (2011). Reticulocyte and erythro-

cyte binding-like proteins function cooperatively in invasion of human erythrocytes by malaria parasites. *Infection and Immunity*, 79(3), 1107–1117.

[203] Lopez-Barragan, M. J., Lemieux, J., Quinones, M., Williamson, K. C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K., & Su, X. Z. (2011). Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics*, 12, 587.

[204] Loy, D. E., Liu, W., Li, Y., Learn, G. H., Plenderleith, L. J., Sundararaman, S. A., Sharp, P. M., & Hahn, B. H. (2017). Out of africa: origins and evolution of the human malaria parasites plasmodium falciparum and plasmodium vivax. *International journal for parasitology*, 47(2-3), 87–97.

[205] Lunter, G. & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Research*, 21(6), 936–939.

[206] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D. W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., & Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1, 18–18.

[207] Maier, A. G., Duraisingh, M. T., Reeder, J. C., Patel, S. S., Kazura, J. W., Zimmerman, P. A., & Cowman, A. F. (2003). *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nature medicine*, 9(1), 87–92.

[208] MalariaGEN *Plasmodium falciparum* Community Project (2016). Genomic epidemiology of artemisinin resistant malaria. *Elife*, 5, e08714.

[209] Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O'Brien, J., Djimde, A., Doumbo, O., Zongo, I., Ouedraogo, J. B., Michon, P., Mueller, I., Siba, P., Nzila, A., Borrmann, S., Kiara, S. M., Marsh, K., Jiang, H., Su, X. Z., Amaratunga, C., Fairhurst, R., Socheat, D., Nosten, F., Imwong, M., White, N. J., Sanders, M., Anastasi, E., Alcock, D., Drury, E., Oyola, S., Quail, M. A., Turner, D. J., Ruano-Rubio, V., Jyothi, D., Amenga-Etego, L., Hubbart, C., Jeffreys, A., Rowlands, K., Sutherland, C., Roper, C., Mangano, V., Modiano, D., Tan, J. C., Ferdig, M. T., Amambua-Ngwa, A., Conway, D. J., Takala-Harrison, S., Plowe, C. V., Rayner, J. C., Rockett, K. A., Clark, T. G., Newbold, C. I., Berriman, M., MacInnis, B., & Kwiatkowski, D. P. (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*, 487(7407), 375–379.

[210] Markus, M. B. (2012). Dormancy in mammalian malaria. *Trends in Parasitology*, 28(2), 39 – 45.

[211] Maxmen, A. (2016). Back on TRAC: New trial launched in bid to outpace multidrug-resistant malaria. *Nat Med*, 22(3), 220–1.

[212] Mayer, D. C. G., Cofie, J., Jiang, L., Hartl, D. L., Tracy, E., Kabat, J., Mendoza, L. H., & Miller, L. H. (2009). Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proc Natl Acad Sci U S A*, 106(13), 5348–5352.

[213] Mbengue, A., Bhattacharjee, S., Pandharkar, T., Liu, H., Estiu, G., Stahelin, R. V., Rizk, S., Njimoh, D. L., Ryan, Y., Chotivanich, K., Nguon, C., Ghorbal, M., Lopez-Rubio, J.-J., Pfrender, M., Emrich, S., Mohandas, N., Dondorp, A. M., Wiest, O., & Haldar, K. (2015). A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. *Nature*, 520(7549), 683–687.

[214] McFadden, G. I., Reith, M. E., Munholland, J., & Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature*, 381, 482 EP –.

[215] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9), 1297–303.

[216] Mellars, P. (2006). Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science*, 313(5788), 796–800.

[217] Ménard, D., Barnadas, C., Bouchier, C., Henry-Halldin, C., Gray, L. R., Ratsimbasoa, A., Thonier, V., Carod, J.-F., Domarle, O., Colin, Y., Bertrand, O., Picot, J., King, C. L., Grimberg, B. T., Mercereau-Puijalon, O., & Zimmerman, P. A. (2010). *Plasmodium vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc Natl Acad Sci U S A*, 107(13), 5967–5971.

[218] Menard, D., Chan, E. R., Benedet, C., Ratsimbasoa, A., Kim, S., Chim, P., Do, C., Witkowski, B., Durand, R., Thellier, M., Severini, C., Legrand, E., Musset, L., Nour, B. Y., Mercereau-Puijalon, O., Serre, D., & Zimmerman, P. A. (2013). Whole genome sequencing of field isolates reveals a common duplication of the Duffy binding protein gene in Malagasy *Plasmodium vivax* strains. *PLoS Negl Trop Dis*, 7(11), e2489.

[219] Menard, D., Khim, N., Beghain, J., Adegnika, A. A., Shafiul-Alam, M., Amodu, O., Rahim-Awab, G., Barnadas, C., Berry, A., Boum, Y., Bustos, M. D., Cao, J., Chen, J. H., Collet, L., Cui, L., Thakur, G. D., Dieye, A., Djalle, D., Dorkenoo, M. A., Eboumbou-Moukoko,

C. E., Espino, F. E., Fandeur, T., Ferreira-da Cruz, M. F., Fola, A. A., Fuehrer, H. P., Hassan, A. M., Herrera, S., Hongvanthong, B., Houze, S., Ibrahim, M. L., Jahirul-Karim, M., Jiang, L., Kano, S., Ali-Khan, W., Khanthavong, M., Kremsner, P. G., Lacerda, M., Leang, R., Leelawong, M., Li, M., Lin, K., Mazarati, J. B., Menard, S., Morlais, I., Muhindo-Mavoko, H., Musset, L., Na-Bangchang, K., Nambozi, M., Niare, K., Noedl, H., Ouedraogo, J. B., Pillai, D. R., Pradines, B., Quang-Phuc, B., Ramharter, M., Randrianarivelojosia, M., Sattabongkot, J., Sheikh-Omar, A., Silue, K. D., Sirima, S. B., Sutherland, C., Syafruddin, D., Tahar, R., Tang, L. H., Toure, O. A., Tshibangu-wa Tshibangu, P., Vigan-Womas, I., Warsame, M., Wini, L., Zakeri, S., Kim, S., Eam, R., Berne, L., Khean, C., Chy, S., Ken, M., Loch, K., Canier, L., Duru, V., Legrand, E., Barale, J. C., Stokes, B., Straimer, J., Witkowski, B., Fidock, D. A., Rogier, C., Ringwald, P., Ariey, F., Mercereau-Puijalon, O., & Consortium, K. (2016). A worldwide map of *Plasmodium falciparum* K13-propeller polymorphisms. *N Engl J Med*, 374(25), 2453–64.

[220] Miller, L. H., Mason, S. J., Clyde, D. F., & McGinniss, M. H. (1976). The resistance factor to *Plasmodium vivax* in blacks. *New England Journal of Medicine*, 295(6), 302–304. PMID: 778616.

[221] Miotto, O., Amato, R., Ashley, E. A., MacInnis, B., Almagro-Garcia, J., Amaratunga, C., Lim, P., Mead, D., Oyola, S. O., Dhorda, M., Imwong, M., Woodrow, C., Manske, M., Stalker, J., Drury, E., Campino, S., Amenga-Etego, L., Thanh, T. N., Tran, H. T., Ringwald, P., Bethell, D., Nosten, F., Phyo, A. P., Pukrittayakamee, S., Chotivanich, K., Chuor, C. M., Nguon, C., Suon, S., Sreng, S., Newton, P. N., Mayxay, M., Khanthavong, M., Hongvanthong, B., Htut, Y., Han, K. T., Kyaw, M. P., Faiz, M. A., Fanello, C. I., Onyamboko, M., Mokuolu, O. A., Jacob, C. G., Takala-Harrison, S., Plowe, C. V., Day, N. P., Dondorp, A. M., Spencer, C. C., McVean, G., Fairhurst, R. M., White, N. J., & Kwiatkowski, D. P. (2015). Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. *Nat Genet*, 47(3), 226–34.

[222] Molineaux, L., Storey, J., Cohen, J. E., & Thomas, A. (1980). A longitudinal study of human malaria in the West African Savanna in the absence of control measures: relationships between different *Plasmodium* species, in particular *P. falciparum* and *P. malariae*. *Am J Trop Med Hyg*, 29(5), 725–37.

[223] Mombo-Ngoma, G., Kleine, C., Basra, A., Wurbel, H., Diop, D. A., Capan, M., Adegnika, A. A., Kurth, F., Mordmuller, B., Joanny, F., Kremsner, P. G., Ramharter, M., & Belard, S. (2012). Prospective evaluation of artemether-lumefantrine for the treatment of non-falciparum and mixed-species malaria in Gabon. *Malar J*, 11, 120.

[224] Moon, R. W., Hall, J., Rangkuti, F., Ho, Y. S., Almond, N., Mitchell, G. H., Pain, A., Holder, A. A., & Blackman, M. J. (2013). Adaptation of the genetically tractable malaria pathogen

Plasmodium knowlesi to continuous culture in human erythrocytes. *Proc Natl Acad Sci U S A*, 110(2), 531–536.

[225] Moon, R. W., Sharaf, H., Hastings, C. H., Ho, Y. S., Nair, M. B., Rchiad, Z., Knuepfer, E., Ramaprasad, A., Mohring, F., Amir, A., Yusuf, N. A., Hall, J., Almond, N., Lau, Y. L., Pain, A., Blackman, M. J., & Holder, A. A. (2016). Normocyte-binding protein required for human erythrocyte invasion by the zoonotic malaria parasite *Plasmodium knowlesi*. *Proc Natl Acad Sci U S A*, 113(26), 7231–7236.

[226] Moore, K. A., Simpson, J. A., Wiladphaingern, J., Min, A. M., Pimanpanarak, M., Paw, M. K., Raksuansak, J., Pukrittayakamee, S., Fowkes, F. J. I., White, N. J., Nosten, F., & McGready, R. (2017). Influence of the number and timing of malaria episodes during pregnancy on prematurity and small-for-gestational-age in an area of low transmission. *BMC Medicine*, 15(1), 117.

[227] Morgulis, A., Gertz, E. M., Schaffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol*, 13(5), 1028–40.

[228] Mu, J., Awadalla, P., Duan, J., McGee, K. M., Keebler, J., Seydel, K., McVean, G. A., & Su, X. Z. (2007). Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nat Genet*, 39(1), 126–30.

[229] Mu, J., Joy, D. A., Duan, J., Huang, Y., Carlton, J., Walker, J., Barnwell, J., Beerli, P., Charleston, M. A., Pybus, O. G., & Su, X.-z. (2005). Host switch leads to emergence of *Plasmodium vivax* malaria in humans. *Molecular Biology and Evolution*, 22(8), 1686–1693.

[230] Mu, J., Myers, R. A., Jiang, H., Liu, S., Ricklefs, S., Waisberg, M., Chotivanich, K., Wilairata, P., Krudsood, S., White, N. J., Udomsangpetch, R., Cui, L., Ho, M., Ou, F., Li, H., Song, J., Li, G., Wang, X., Seila, S., Sokunthea, S., Socheat, D., Sturdevant, D. E., Porcella, S. F., Fairhurst, R. M., Wellems, T. E., Awadalla, P., & Su, X.-z. (2010). *Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature genetics*, 42(3), 268–271.

[231] Mueller, I., Galinski, M. R., Baird, J. K., Carlton, J. M., Kochar, D. K., Alonso, P. L., & del Portillo, H. A. (2009). Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. *The Lancet Infectious Diseases*, 9(9), 555 – 566.

[232] Murphy, S. & Breman, J. (2001). Gaps in the childhood malaria burden in Africa: cerebral malaria, neurological sequelae, anemia, respiratory distress, hypoglycemia, and complications of pregnancy. *The American Journal of Tropical Medicine and Hygiene*, 64(1 suppl), 57 – 67.

[233] Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2), 275–290.

[234] Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, 21(suppl$_2$), *ii*79 – –*ii*85.

[235] Nabarro, D. N. & Tayler, E. M. (1998). The roll back malaria campaign. *Science*, 280(5372), 2067.

[236] Nadalin, F., Vezzi, F., & Policriti, A. (2012). GapFiller: a *de novo* assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, 13 Suppl 14, S8.

[237] Nájera, J., González-Silva, M., & Alonso, P. L. (2011). Some lessons for the future from the global malaria eradication programme (1955–1969). *PLoS Medicine*, 8(1), e1000412.

[238] Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., & Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, 43(Database issue), D130–7.

[239] Neafsey, D. E., Galinsky, K., Jiang, R. H. Y., Young, L., Sykes, S. M., Saif, S., Gujja, S., Goldberg, J. M., Young, S., Zeng, Q., Chapman, S. B., Dash, A. P., Anvikar, A. R., Sutton, P. L., Birren, B. W., Escalante, A. A., Barnwell, J. W., & Carlton, J. M. (2012). The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. *Nature genetics*, 44(9), 1046–1050.

[240] Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5), 418–26.

[241] Nekrutenko, A., Makova, K. D., & Li, W. H. (2002). The k(a)/k(s) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res*, 12(1), 198–202.

[242] Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12, 443 EP –.

[243] Noedl, H., Se, Y., Schaecher, K., Smith, B. L., Socheat, D., Fukuda, M. M., & Artemisinin Resistance in Cambodia 1 Study, C. (2008). Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med*, 359(24), 2619–20.

[244] Nolder, D., Oguike, M. C., Maxwell-Scott, H., Niyazi, H. A., Smith, V., Chiodini, P. L., & Sutherland, C. J. (2013). An observational study of malaria in British travellers: *Plasmodium ovale wallikeri* and *Plasmodium ovale curtisi* differ significantly in the duration of latency. *BMJ Open*, 3(5).

[245] Nosten, F., ter Kuile, F., Chongsuphajaisiddhi, T., Luxemburger, C., Webster, H. K., Edstein, M., Phaipun, L., Thew, K. L., & White, N. J. (1991). Mefloquine-resistant falciparum malaria on the Thai-Burmese border. *Lancet*, 337(8750), 1140–3.

[246] Nosten, F., van Vugt, M., Price, R., Luxemburger, C., Thway, K. L., Brockman, A., McGready, R., ter Kuile, F., Looareesuwan, S., & White, N. J. (2000). Effects of artesunate-mefloquine combination on incidence of *Plasmodium falciparum* malaria and mefloquine resistance in western Thailand: a prospective study. *Lancet*, 356(9226), 297–302.

[247] Nunes-Alves, C. (2015). RIFINs promote rosette formation during malaria. *Nature Reviews Microbiology*, 13, 250 EP –.

[248] Ochoa, D. & Pazos, F. (2010). Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, 26(10), 1370–1.

[249] Oguike, M. C., Betson, M., Burke, M., Nolder, D., Stothard, J. R., Kleinschmidt, I., Proietti, C., Bousema, T., Ndounga, M., Tanabe, K., Ntege, E., Culleton, R., & Sutherland, C. J. (2011). *Plasmodium ovale curtisi* and *Plasmodium ovale wallikeri* circulate simultaneously in African communities. *International Journal for Parasitology*, 41(6), 677 – 683.

[250] Oguike, M. C. & Sutherland, C. J. (2015). Dimorphism in genes encoding sexual-stage proteins of *Plasmodium ovale curtisi* and *Plasmodium ovale wallikeri*. *International Journal for Parasitology*, 45(7), 449–454.

[251] Okombo, J., Abdi, A. I., Kiara, S. M., Mwai, L., Pole, L., Sutherland, C. J., Nzila, A., & Ochola-Oyier, L. I. (2013). Repeat polymorphisms in the low-complexity regions of *Plasmodium falciparum* ABC transporters and associations with *in vitro* antimalarial responses. *Antimicrob Agents Chemother*, 57(12), 6196–204.

[252] Olivieri, A., Camarda, G., Bertuccini, L., van de Vegte-Bolmer, M., Luty, A. J., Sauerwein, R., & Alano, P. (2009). The *Plasmodium falciparum* protein Pfg27 is dispensable for gametocyte and gamete production, but contributes to cell integrity during gametocytogenesis. *Mol Microbiol*, 73(2), 180–93.

[253] Ollomo, B., Durand, P., Prugnolle, F., Douzery, E., Arnathau, C., Nkoghe, D., Leroy, E., & Renaud, F. (2009). A new malaria agent in African hominids. *PLoS Pathog*, 5(5), e1000446.

[254] Orjuela-Sánchez, P., de Santana Filho, F. S., Machado-Lima, A., Chehuan, Y. F., Costa, M. R. F., Alecrim, M. d. G. C., & del Portillo, H. A. (2009). Analysis of single-nucleotide polymorphisms in the *crt-o* and *mdr1* genes of *Plasmodium vivax* among chloroquine-resistant isolates from the Brazilian Amazon region. *Antimicrobial Agents and Chemotherapy*, 53(8), 3561–3564.

[255] Otto, T., Böhme, U., Sanders, M., Reid, A., Bruske, E., Duffy, C., Bull, P., Pearson, R., Abdi, A., Dimonte, S., Stewart, L., Campino, S., Kekre, M., Hamilton, W., Claessens, A., Volkman, S., Ndiaye,

D., Amambua-Ngwa, A., Diakite, M., Fairhurst, R., Conway, D., Franck, M., Newbold, C., & Berriman, M. (2018a). Long read assemblies of geographically dispersed *Plasmodium falciparum* isolates reveal highly structured subtelomeres. *Wellcome Open Research*, 3(52).

[256] Otto, T. D., Bohme, U., Jackson, A. P., Hunt, M., Franke-Fayard, B., Hoeijmakers, W. A., Religa, A. A., Robertson, L., Sanders, M., Ogun, S. A., Cunningham, D., Erhart, A., Billker, O., Khan, S. M., Stunnenberg, H. G., Langhorne, J., Holder, A. A., Waters, A. P., Newbold, C. I., Pain, A., Berriman, M., & Janse, C. J. (2014a). A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol*, 12, 86.

[257] Otto, T. D., Dillon, G. P., Degrave, W. S., & Berriman, M. (2011). RATT: Rapid annotation transfer tool. *Nucleic Acids Res*, 39(9), e57.

[258] Otto, T. D., Gilabert, A., Crellen, T., Böhme, U., Arnathau, C., Sanders, M., Oyola, S. O., Okouga, A. P., Boundenga, L., Willaume, E., Ngoubangoye, B., Moukodoum, N. D., Paupy, C., Durand, P., Rougeron, V., Ollomo, B., Renaud, F., Newbold, C., Berriman, M., & Prugnolle, F. (2018b). Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria. *Nature Microbiology*, 3(6), 687–697.

[259] Otto, T. D., Rayner, J. C., Bohme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A. W., Prugnolle, F., Conway, D. J., Newbold, C., & Berriman, M. (2014b). Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nat Commun*, 5, 4754.

[260] Otto, T. D., Sanders, M., Berriman, M., & Newbold, C. (2010). Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, 26(14), 1704–7.

[261] Oyola, S. O., Ariani, C. V., Hamilton, W. L., Kekre, M., Amenga-Etego, L. N., Ghansah, A., Rutledge, G. G., Redmond, S., Manske, M., Jyothi, D., Jacob, C. G., Otto, T. D., Rockett, K., Newbold, C. I., Berriman, M., & Kwiatkowski, D. P. (2016). Whole genome sequencing of *Plasmodium falciparum* from dried blood spots using selective whole genome amplification. *Malaria Journal*, 15, 597.

[262] Oyola, S. O., Gu, Y., Manske, M., Otto, T. D., O'Brien, J., Alcock, D., MacInnis, B., Berriman, M., Newbold, C. I., Kwiatkowski, D. P., Swerdlow, H. P., & Quail, M. A. (2013). Efficient depletion of host DNA contamination in malaria clinical sequencing. *Journal of Clinical Microbiology*, 51(3), 745–751.

[263] Oyola, S. O., Manske, M., Campino, S., Claessens, A., Hamilton, W. L., Kekre, M., Drury, E., Mead, D., Gu, Y., Miles, A., MacInnis, B., Newbold, C., Berriman, M., & Kwiatkowski, D. P. (2014). Opti-

mized whole-genome amplification strategy for extremely AT-biased template. *DNA Research*, 21(6), 661–671.

[264] Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D. J., MacInnis, B., Kwiatkowski, D. P., Swerdlow, H. P., & Quail, M. A. (2012). Optimizing illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics*, 13(1), 1.

[265] Padley, D., Moody, A. H., Chiodini, P. L., & Saldanha, J. (2003). Use of a rapid, single-round, multiplex PCR to detect malarial parasites and identify the species present. *Ann Trop Med Parasitol*, 97(2), 131–7.

[266] Pain, A., Bohme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., Mourier, T., Mistry, J., Pasini, E. M., Aslett, M. A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T. J., Cherevach, I., Chillingworth, T., Clark, T. G., Galinski, M. R., Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C. S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I. M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T. J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A. R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M. A., Cowman, A. F., Turner, C. M., Rajandream, M. A., Kocken, C. H., Thomas, A. W., Newbold, C. I., Barrell, B. G., & Berriman, M. (2008). The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature*, 455(7214), 799–803.

[267] Park, D. J., Lukens, A. K., Neafsey, D. E., Schaffner, S. F., Chang, H.-H., Valim, C., Ribacke, U., Van Tyne, D., Galinsky, K., Galligan, M., Becker, J. S., Ndiaye, D., Mboup, S., Wiegand, R. C., Hartl, D. L., Sabeti, P. C., Wirth, D. F., & Volkman, S. K. (2012). Sequence-based association and selection scans identify drug resistance loci in the *Plasmodium falciparum* malaria parasite. *Proc Natl Acad Sci USA*, 109(32), 13052–13057.

[268] Parobek, C. M., Parr, J. B., Brazeau, N. F., Lon, C., Chaorattanakawee, S., Gosi, P., Barnett, E. J., Norris, L. D., Meshnick, S. R., Spring, M. D., Lanteri, C. A., Bailey, J. A., Saunders, D. L., Lin, J. T., & Juliano, J. J. (2017). Partner-drug resistance and population substructuring of artemisinin-resistant *Plasmodium falciparum* in Cambodia. *Genome Biology and Evolution*, 9(6), 1673–1686.

[269] Pasini, E. M., Böhme, U., Rutledge, G. G., Voorberg-Van der Wel, A., Sanders, M., Berriman, M., Kocken, C. H., & Otto, T. D. (2017). An improved *Plasmodium cynomolgi* genome assembly reveals an unexpected methyltransferase gene expansion. *Wellcome Open Research*, 2, 42.

[270] Pasloske, B. L. & Howard, R. J. (1994). Malaria, the red cell, and the endothelium. *Annual Review of Medicine*, 45(1), 283–295.

[271] Pastoor, D. (2015). Extending Rstudio's functionality to accelerate modeler workflows via shiny applications. *Journal of Pharmacokinetics and Pharmacodynamics*, 42, S68–S69.

[272] Payne, D. (1987). Spread of chloroquine resistance in *Plasmodium falciparum*. *Parasitology Today*, 3(8), 241–246.

[273] Payne, R. O., Silk, S. E., Elias, S. C., Miura, K., Diouf, A., Galaway, F., de Graaf, H., Brendish, N. J., Poulton, I. D., Griffiths, O. J., Edwards, N. J., Jin, J., Labbé, G. M., Alanine, D. G., Siani, L., Di Marco, S., Roberts, R., Green, N., Berrie, E., Ishizuka, A. S., Nielsen, C. M., Bardelli, M., Partey, F. D., Ofori, M. F., Barfod, L., Wambua, J., Murungi, L. M., Osier, F. H., Biswas, S., McCarthy, J. S., Minassian, A. M., Ashfield, R., Viebig, N. K., Nugent, F. L., Douglas, A. D., Vekemans, J., Wright, G. J., Faust, S. N., Hill, A. V., Long, C. A., Lawrie, A. M., & Draper, S. J. (2017). Human vaccination against rh5 induces neutralizing antimalarial antibodies that inhibit rh5 invasion complex interactions. *JCI Insight*, 2(21), e96381.

[274] Pearson, R. D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C., Suon, S., Mao, S., Noviyanti, R., Trimarsanto, H., Marfurt, J., Anstey, N. M., William, T., Boni, M. F., Dolecek, C., Tran, H. T., White, N. J., Michon, P., Siba, P., Tavul, L., Harrison, G., Barry, A., Mueller, I., Ferreira, M. U., Karunaweera, N., Randrianarivelojosia, M., Gao, Q., Hubbart, C., Hart, L., Jeffery, B., Drury, E., Mead, D., Kekre, M., Campino, S., Manske, M., Cornelius, V. J., MacInnis, B., Rockett, K. A., Miles, A., Rayner, J. C., Fairhurst, R. M., Nosten, F., Price, R. N., & Kwiatkowski, D. P. (2016). Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nature Genetics*, 48, 959 EP –.

[275] Penman, B., Buckee, C., Gupta, S., & Nee, S. (2010). Genome-wide association studies in *Plasmodium* species. *BMC Biology*, 8, 90–90.

[276] Peterson, D. S., Milhous, W. K., & Wellems, T. E. (1990). Molecular basis of differential resistance to cycloguanil and pyrimethamine in *Plasmodium falciparum* malaria. *Proc Natl Acad Sci U S A*, 87(8), 3018–3022.

[277] Peterson, D. S., Walliker, D., & Wellems, T. E. (1988). Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine in falciparum malaria. *Proc Natl Acad Sci U S A*, 85(23), 9114–9118.

[278] Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*, 98(17), 9748–9753.

[279] Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, 27(5), xi–xiii.

[280] Pickard, A. L., Wongsrichanalai, C., Purfield, A., Kamwendo, D., Emery, K., Zalewski, C., Kawamoto, F., Miller, R. S., & Meshnick, S. R. (2003). Resistance to antimalarials in southeast Asia and genetic polymorphisms in *pfmdr1*. *Antimicrobial Agents and Chemotherapy*, 47(8), 2418–2423.

[281] Pinheiro, M. M., Ahmed, M. A., Millar, S. B., Sanderson, T., Otto, T. D., Lu, W. C., Krishna, S., Rayner, J. C., & Cox-Singh, J. (2015). *Plasmodium knowlesi* genome sequences from clinical isolates reveal extensive genomic dimorphism. *PLoS One*, 10(4), 1–16.

[282] Price, R. N., Uhlemann, A. C., Brockman, A., McGready, R., Ashley, E., Phaipun, L., Patel, R., Laing, K., Looareesuwan, S., White, N. J., Nosten, F., & Krishna, S. (2004). Mefloquine resistance in *Plasmodium falciparum* and increased *pfmdr1* gene copy number. *Lancet*, 364(9432), 438–447.

[283] Price, R. N., Uhlemann, A. C., van Vugt, M., Brockman, A., Hutagalung, R., Nair, S., Nash, D., Singhasivanon, P., Anderson, T. J., Krishna, S., White, N. J., & Nosten, F. (2006). Molecular and pharmacological determinants of the therapeutic response to artemether-lumefantrine in multidrug-resistant *Plasmodium falciparum* malaria. *Clin Infect Dis*, 42(11), 1570–7.

[284] Protopopoff, N., Mosha, J. F., Lukole, E., Charlwood, J. D., Wright, A., Mwalimu, C. D., Manjurano, A., Mosha, F. W., Kisinza, W., Kleinschmidt, I., & Rowland, M. (2018). Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial. *The Lancet*, 391(10130), 1577–1588.

[285] Prudêncio, M., Rodriguez, A., & Mota, M. M. (2006). The silent path to thousands of merozoites: the *Plasmodium* liver stage. *Nature Reviews Microbiology*, 4, 849 EP –.

[286] Prugnolle, F., Rougeron, V., Becquart, P., Berry, A., Makanga, B., Rahola, N., Arnathau, C., Ngoubangoye, B., Menard, S., Willaume, E., Ayala, F. J., Fontenille, D., Ollomo, B., Durand, P., Paupy, C., & Renaud, F. (2013). Diversity, host switching and evolution of *Plasmodium vivax* infecting African great apes. *Proc Natl Acad Sci U S A*, 110(20), 8123–8128.

[287] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575.

[288] Raj, D. K., Mu, J., Jiang, H., Kabat, J., Singh, S., Sullivan, M., Fay, M. P., McCutchan, T. F., & Su, X.-z. (2009). Disruption of a *Plasmodium falciparum* multidrug resistance-associated protein (PfMRP) alters its fitness and transport of antimalarial drugs and glutathione. *The Journal of Biological Chemistry*, 284(12), 7687–7696.

[289] Rajahram, G. S., Barber, B. E., William, T., Grigg, M. J., Menon, J., Yeo, T. W., & Anstey, N. M. (2016). Falling *Plasmodium knowlesi* malaria death rate among adults despite rising incidence, Sabah, Malaysia, 2010–2014. *Emerging Infectious Diseases*, 22(1), 41–48.

[290] Ratcliff, A., Siswantoro, H., Kenangalem, E., Maristela, R., Wuwung, R. M., Laihad, F., Ebsworth, E. P., Anstey, N. M., Tjitra, E., & Price, R. N. (2007). Two fixed-dose artemisinin combinations for drug-resistant falciparum and vivax malaria in Papua, Indonesia: an open-label randomised comparison. *Lancet*, 369(9563), 757–65.

[291] Rayner, J. C., Liu, W., Peeters, M., Sharp, P. M., & Hahn, B. H. (2011). A plethora of *Plasmodium* species in wild apes: a source of human infection? *Trends in Parasitology*, 27(5), 222–229.

[292] Reid, A. J. (2015). Large, rapidly evolving gene families are at the forefront of host–parasite interactions in apicomplexa. *Parasitology*, 142(S1), S57–S70.

[293] Rich, S. M. & Xu, G. (2011). Resolving the phylogeny of malaria parasites. *Proc Natl Acad Sci U S A*, 108(32), 12973–12974.

[294] Roberts, A. & Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*, 10(1), 71–3.

[295] Roques, M., Wall, R. J., Douglass, A. P., Ramaprasad, A., Ferguson, D. J., Kaindama, M. L., Brusini, L., Joshi, N., Rchiad, Z., Brady, D., Guttery, D. S., Wheatley, S. P., Yamano, H., Holder, A. A., Pain, A., Wickstead, B., & Tewari, R. (2015). *Plasmodium* P-type cyclin CYC3 modulates endomitotic growth during oocyst development in mosquitoes. *PLoS Pathog*, 11(11), e1005273.

[296] Rosenberg, R. & Rungsiwongse, J. (1991). The number of sporozoites produced by individual malaria oocysts. *The American Journal of Tropical Medicine and Hygiene*, 45(5), 574–577.

[297] Rossi, G., De Smet, M., Khim, N., Kindermans, J. M., & Menard, D. (2017). Emergence of *Plasmodium falciparum* triple mutant in Cambodia. *Lancet Infect Dis*, 17(12), 1233.

[298] Roucher, C., Rogier, C., Sokhna, C., Tall, A., & Trape, J. F. (2014). A 20-year longitudinal study of *Plasmodium ovale* and *Plasmodium malariae* prevalence and morbidity in a West African population. *PLoS One*, 9(2), e87169.

[299] Rowe, J. A., Moulds, J. M., Newbold, C. I., & Miller, L. H. (1997). *P. falciparum* rosetting mediated by a parasite-variant erythrocyte membrane protein and complement-receptor 1. *Nature*, 388, 292 EP –.

[300] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10), 944–5.

[301] Rutledge, G. G., Bohme, U., Sanders, M., Reid, A. J., Cotton, J. A., Maiga-Ascofare, O., Djimde, A. A., Apinjoh, T. O., Amenga-Etego, L., Manske, M., Barnwell, J. W., Renaud, F., Ollomo, B.,

Prugnolle, F., Anstey, N. M., Auburn, S., Price, R. N., McCarthy, J. S., Kwiatkowski, D. P., Newbold, C. I., Berriman, M., & Otto, T. D. (2017). *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. *Nature*, 542(7639), 101–104.

[302] Rutledge, G. G. & Otto, T. D. (2016). Last parasite standing. *Nat Rev Microbiol*, 15(1), 4.

[303] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909), 832–7.

[304] Salanti, A., Dahlbäck, M., Turner, L., Nielsen, M. A., Barfod, L., Magistrado, P., Jensen, A. T., Lavstsen, T., Ofori, M. F., Marsh, K., Hviid, L., & Theander, T. G. (2004). Evidence for the involvement of VAR2CSA in pregnancy-associated malaria. *Journal of Experimental Medicine*, 200(9), 1197–1203.

[305] Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12), 5463–5467.

[306] Saul, A. (1999). The role of variant surface antigens on malaria-infected red blood cells. *Parasitology Today*, 15(11), 455–457.

[307] Saunders, D. L., Vanachayangkul, P., Lon, C., & Program, U. (2014). Dihydroartemisinin-piperaquine failure in Cambodia. *New England Journal of Medicine*, 371(5), 484–485.

[308] Schaer, J., Perkins, S. L., Decher, J., Leendertz, F. H., Fahr, J., Weber, N., & Matuschewski, K. (2013). High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc Natl Acad Sci U S A*, 110(43), 17415–17419.

[309] Scherf, A., Figueiredo, L. M., & Freitas-Junior, L. H. (2001). *Plasmodium* telomeres: a pathogen's perspective. *Current Opinion in Microbiology*, 4(4), 409 – 414.

[310] Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J., & Lanzer, M. (1998). Antigenic variation in malaria: *in situ* switching, relaxed and mutually exclusive transcription of *var* genes during intra-erythrocytic development in *Plasmodium falciparum*. *The EMBO Journal*, 17(18), 5418–5426.

[311] Scopel, K. K., Fontes, C. J., Nunes, A. C., Horta, M. F., & Braga, E. M. (2004). High prevalence of *Plamodium malariae* infections in a Brazilian Amazon endemic area (Apiacas-Mato Grosso State) as detected by polymerase chain reaction. *Acta Trop*, 90(1), 61–4.

[312] Shanks, G. D. & White, N. J. (2013). The activation of vivax malaria hypnozoites by infectious diseases. *The Lancet Infectious Diseases*, 13(10), 900 – 906.

[313] Shaw-Saliba, K., Thomson-Luque, R., Obaldía, Nicanor, I., Nuñez, M., Dutary, S., Lim, C., Barnes, S., Kocken, C. H. M., Duraisingh, M. T., Adams, J. H., & Pasini, E. M. (2016). Insights into an optimization of *Plasmodium vivax* Sal-1 *in vitro* culture: The Aotus primate model. *PLoS Neglected Tropical Diseases*, 10(7), e0004870–.

[314] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550, 345 EP –.

[315] Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26, 1135 EP –.

[316] Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741), 1728–1732.

[317] Shortt, H. E. & Garnham, P. C. (2000). Demonstration of a persisting exo-erythrocytic cycle in *Plasmodium cynomolgi* and its bearing on the production of relapses. *Bulletin of the World Health Organization*, 78(12), 1447–1449.

[318] Siala, E., Khalfaoui, M., Bouratbine, A., Hamdi, S., Hili, K., & Aoun, K. (2005). Relapse of *Plasmodium malariae* malaria 20 years after living in an endemic area. *Presse Med*, 34(5), 371–2.

[319] Sidhu, A. B. S., Uhlemann, A.-C., Valderramos, S. G., Valderramos, J.-C., Krishna, S., & Fidock, D. A. (2006). Decreasing *pfmdr1* copy number in *Plasmodium falciparum* malaria heightens susceptibility to mefloquine, lumefantrine, halofantrine, quinine, and artemisinin. *The Journal of infectious diseases*, 194(4), 528–535.

[320] Sidhu, A. B. S., Valderramos, S. G., & Fidock, D. A. (2005). *pfmdr1* mutations contribute to quinine resistance and enhance mefloquine and artemisinin sensitivity in *Plasmodium falciparum*. *Molecular Microbiology*, 57(4), 913–926.

[321] Sidhu, A. B. S., Verdier-Pinard, D., & Fidock, D. A. (2002). Chloroquine resistance in *Plasmodium falciparum* malaria parasites conferred by *pfcrt* mutations. *Science*, 298(5591), 210–213.

[322] Silva, J. C., Egan, A., Arze, C., Spouge, J. L., & Harris, D. G. (2015). A new method for estimating species age supports the coexistence of malaria parasites and their mammalian hosts. *Mol Biol Evol*, 32(5), 1354–64.

[323] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6), 1117–1123.

[324] Sinden, R. E. (1999). *Plasmodium* differentiation in the mosquito. *Parassitologia*, 41(1-3), 139–148.

[325] Singh, B., Sung, L. K., Matusop, A., Radhakrishnan, A., Shamsul, S. S., Cox-Singh, J., Thomas, A., & Conway, D. J. (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *The Lancet*, 363(9414), 1017 – 1024.

[326] Siwo, G. H., Tan, A., Button-Simons, K. A., Samarakoon, U., Checkley, L. A., Pinapati, R. S., & Ferdig, M. T. (2015). Predicting functional and regulatory divergence of a drug resistance transporter gene in the human malaria parasite. *BMC Genomics*, 16, 115.

[327] Smith, A., Denholm, J., Shortt, J., & Spelman, D. (2011). *Plasmodium* species co-infection as a cause of treatment failure. *Travel Med Infect Dis*, 9(6), 306–9.

[328] Snow, R. W. & Marsh, K. (2010). Malaria in Africa: progress and prospects in the decade since the Abuja Declaration. *Lancet*, 376(9735), 137–139.

[329] Spence, P. J., Jarra, W., Lévy, P., Reid, A. J., Chappell, L., Brugat, T., Sanders, M., Berriman, M., & Langhorne, J. (2013). Vector transmission regulates immune control of *Plasmodium* virulence. *Nature*, 498(7453), 228–231.

[330] Spring, M. D., Lin, J. T., Manning, J. E., Vanachayangkul, P., Somethy, S., Bun, R., Se, Y., Chann, S., Ittiverakul, M., Sia-Ngam, P., Kuntawunginn, W., Arsanok, M., Buathong, N., Chaorattanakawee, S., Gosi, P., Ta-Aksorn, W., Chanarat, N., Sundrakes, S., Kong, N., Heng, T. K., Nou, S., Teja-Isavadharm, P., Pichyangkul, S., Phann, S. T., Balasubramanian, S., Juliano, J. J., Meshnick, S. R., Chour, C. M., Prom, S., Lanteri, C. A., Lon, C., & Saunders, D. L. (2015). Dihydroartemisinin-piperaquine failure associated with a triple mutant including kelch13 C580Y in Cambodia: an observational cohort study. *Lancet Infectious Diseases*, 15(6), 683–691.

[331] Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol*, 57(5), 758–71.

[332] Stamatakis, A., Ludwig, T., & Meier, H. (2005). RAxML-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4), 456–63.

[333] Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res*, 34(Web Server issue), W435–9.

[334] Straimer, J., Gnadig, N. F., Witkowski, B., Amaratunga, C., Duru, V., Ramadani, A. P., Dacheux, M., Khim, N., Zhang, L., Lam, S., Gregory, P. D., Urnov, F. D., Mercereau-Puijalon, O., Benoit-Vical, F., Fairhurst, R. M., Menard, D., & Fidock, D. A. (2015). Drug resistance. K13-propeller mutations confer artemisinin resistance in *Plasmodium falciparum* clinical isolates. *Science*, 347(6220), 428–31.

[335] Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., Pollok, J.-M., Menard, R., & Heussler, V. T. (2006). Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science*, 313(5791), 1287–1290.

[336] Su, X., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., Ravetch, J. A., & Wellems, T. E. (1995). The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell*, 82(1), 89 – 100.

[337] Summers, R. L., Dave, A., Dolstra, T. J., Bellanca, S., Marchetti, R. V., Nash, M. N., Richards, S. N., Goh, V., Schenk, R. L., Stein, W. D., Kirk, K., Sanchez, C. P., Lanzer, M., & Martin, R. E. (2014). Diverse mutational pathways converge on saturable chloroquine transport via the malaria parasite's chloroquine resistance transporter. *Proc Natl Acad Sci U S A*, 111(17), E1759–E1767.

[338] Sundararaman, S. A., Plenderleith, L. J., Liu, W., Loy, D. E., Learn, G. H., Li, Y., Shaw, K. S., Ayouba, A., Peeters, M., Speede, S., Shaw, G. M., Bushman, F. D., Brisson, D., Rayner, J. C., Sharp, P. M., & Hahn, B. H. (2016). Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat Commun*, 7, 11078.

[339] Sutherland, C. J., Tanomsing, N., Nolder, D., Oguike, M., Jennison, C., Pukrittayakamee, S., Dolecek, C., Hien, T. T., do Rosario, V. E., Arez, A. P., Pinto, J., Michon, P., Escalante, A. A., Nosten, F., Burke, M., Lee, R., Blaze, M., Otto, T. D., Barnwell, J. W., Pain, A., Williams, J., White, N. J., Day, N. P., Snounou, G., Lockhart, P. J., Chiodini, P. L., Imwong, M., & Polley, S. D. (2010). Two nonrecombining sympatric forms of the human malaria parasite *Plasmodium ovale* occur globally. *J Infect Dis*, 201(10), 1544–50.

[340] Sutton, G. G., Owen, W., Adams, M. D., & Kerlavage, A. R. (1995). TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1), 9–19.

[341] Tachibana, S., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., Arisue, N., Palacpac, N. M., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P. L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J. W., Escalante, A. A., Carlton, J. M., & Tanabe, K. (2012). *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet*, 44(9), 1051–5.

[342] Takala-Harrison, S., Jacob, C. G., Arze, C., Cummings, M. P., Silva, J. C., Dondorp, A. M., Fukuda, M. M., Hien, T. T., Mayxay, M., Noedl, H., Nosten, F., Kyaw, M. P., Nhien, N. T., Imwong, M., Bethell, D., Se, Y., Lon, C., Tyner, S. D., Saunders, D. L., Ariey, F., Mercereau-Puijalon, O., Menard, D., Newton, P. N., Khanthavong, M., Hongvanthong, B., Starzengruber, P., Fuehrer, H. P., Swoboda, P., Khan, W. A., Phyo, A. P., Nyunt, M. M., Nyunt, M. H., Brown, T. S., Adams, M., Pepin,

C. S., Bailey, J., Tan, J. C., Ferdig, M. T., Clark, T. G., Miotto, O., MacInnis, B., Kwiatkowski, D. P., White, N. J., Ringwald, P., & Plowe, C. V. (2015). Independent emergence of artemisinin resistance mutations among *Plasmodium falciparum* in Southeast Asia. *J Infect Dis*, 211(5), 670–9.

[343] Talavera, G. & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, 56(4), 564–77.

[344] Talundzic, E., Ravishankar, S., Nayak, V., Patel, D. S., Olsen, C., Sheth, M., Batra, D., Loparev, V., Vannberg, F. O., Udhayakumar, V., & Barnwell, J. W. (2017). First full draft genome sequence of *Plasmodium brasilianum*. *Genome Announcements*, 5(6), e01566–16.

[345] Tavares, J., Formaglio, P., Thiberge, S., Mordelet, E., Van Rooijen, N., Medvinsky, A., Ménard, R., & Amino, R. (2013). Role of host cell traversal by the malaria sporozoite during liver infection. *The Journal of Experimental Medicine*, 210(5), 905–915.

[346] Thanh, N. V., Thuy-Nhien, N., Tuyen, N. T., Tong, N. T., Nha-Ca, N. T., Dong, L. T., Quang, H. H., Farrar, J., Thwaites, G., White, N. J., Wolbers, M., & Hien, T. T. (2017). Rapid decline in the susceptibility of *Plasmodium falciparum* to dihydroartemisinin-piperaquine in the south of Vietnam. *Malar J*, 16(1), 27.

[347] Tonkin, M. L., Roques, M., Lamarque, M. H., Pugnière, M., Douguet, D., Crawford, J., Lebrun, M., & Boulanger, M. J. (2011). Host cell invasion by Apicomplexan parasites: Insights from the co-structure of AMA1 with a RON2 peptide. *Science*, 333(6041), 463–467.

[348] Trager, W. & Jensen, J. (1976). Human malaria parasites in continuous culture. *Science*, 193(4254), 673–675.

[349] Triglia, T., Menting, J. G. T., Wilson, C., & Cowman, A. F. (1997). Mutations in dihydropteroate synthase are responsible for sulfone and sulfonamide resistance in *Plasmodium falciparum*. *Proc Natl Acad Sci U S A*, 94(25), 13944–13949.

[350] Tsai, I. J., Otto, T. D., & Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol*, 11(4), R41.

[351] Tun, K. M., Imwong, M., Lwin, K. M., Win, A. A., Hlaing, T. M., Hlaing, T., Lin, K., Kyaw, M. P., Plewes, K., Faiz, M. A., Dhorda, M., Cheah, P. Y., Pukrittayakamee, S., Ashley, E. A., Anderson, T. J. C., Nair, S., McDew-White, M., Flegg, J. A., Grist, E. P., Guerin, P., Maude, R. J., Smithuis, F., Dondorp, A. M., Day, N. P. J., Nosten, F., White, N. J., & Woodrow, C. J. (2015). Spread of artemisinin-resistant *Plasmodium falciparum* in Myanmar: a cross-sectional survey of the K13 molecular marker. *Lancet Infectious Diseases*, 15(4), 415–421.

[352] Ukaegbu, U. E., Zhang, X., Heinberg, A. R., Wele, M., Chen, Q., & Deitsch, K. W. (2015). A unique virulence gene occupies a principal position in immune evasion by the malaria parasite *Plasmodium falciparum*. *PLoS Genetics*, 11(5), 1–26.

[353] Van Tyne, D., Park, D. J., Schaffner, S. F., Neafsey, D. E., Angelino, E., Cortese, J. F., Barnes, K. G., Rosen, D. M., Lukens, A. K., Daniels, R. F., Milner, Danny A., J., Johnson, C. A., Shlyakhter, I., Grossman, S. R., Becker, J. S., Yamins, D., Karlsson, E. K., Ndiaye, D., Sarr, O., Mboup, S., Happi, C., Furlotte, N. A., Eskin, E., Kang, H. M., Hartl, D. L., Birren, B. W., Wiegand, R. C., Lander, E. S., Wirth, D. F., Volkman, S. K., & Sabeti, P. C. (2011). Identification and functional validation of the novel antimalarial resistance locus PF10_0355 in *Plasmodium falciparum*. *PLoS Genetics*, 7(4), e1001383–.

[354] Veiga, M. I., Dhingra, S. K., Henrich, P. P., Straimer, J., Gnädig, N., Uhlemann, A.-C., Martin, R. E., Lehane, A. M., & Fidock, D. A. (2016). Globally prevalent PfMDR1 mutations modulate *Plasmodium falciparum* susceptibility to artemisinin-based combination therapies. *Nature Communications*, 7, 11553 EP –.

[355] Venkatesan, M., Amaratunga, C., Campino, S., Auburn, S., Koch, O., Lim, P., Uk, S., Socheat, D., Kwiatkowski, D. P., Fairhurst, R. M., & Plowe, C. V. (2012). Using CF11 cellulose columns to inexpensively and effectively remove human DNA from *Plasmodium falciparum*-infected whole blood samples. *Malaria Journal*, 11(1), 41.

[356] Venkatesan, M., Gadalla, N. B., Stepniewska, K., Dahal, P., Nsanzabana, C., Moriera, C., Price, R. N., Martensson, A., Rosenthal, P. J., Dorsey, G., Sutherland, C. J., Guerin, P., Davis, T. M., Menard, D., Adam, I., Ademowo, G., Arze, C., Baliraine, F. N., Berens-Riha, N., Bjorkman, A., Borrmann, S., Checchi, F., Desai, M., Dhorda, M., Djimde, A. A., El-Sayed, B. B., Eshetu, T., Eyase, F., Falade, C., Faucher, J. F., Froberg, G., Grivoyannis, A., Hamour, S., Houze, S., Johnson, J., Kamugisha, E., Kariuki, S., Kiechel, J. R., Kironde, F., Kofoed, P. E., LeBras, J., Malmberg, M., Mwai, L., Ngasala, B., Nosten, F., Nsobya, S. L., Nzila, A., Oguike, M., Otienoburu, S. D., Ogutu, B., Ouedraogo, J. B., Piola, P., Rombo, L., Schramm, B., Some, A. F., Thwing, J., Ursing, J., Wong, R. P., Zeynudin, A., Zongo, I., Plowe, C. V., Sibley, C. H., Group, A. M. M. S., & Wwarn, A. L. (2014). Polymorphisms in *Plasmodium falciparum* chloroquine resistance transporter and multidrug resistance 1 genes: parasite risk factors that affect treatment outcomes for *P. falciparum* malaria after artemether-lumefantrine and artesunate-amodiaquine. *Am J Trop Med Hyg*, 91(4), 833–43.

[357] Vinetz, J. M., Li, J., McCutchan, T. F., & Kaslow, D. C. (1998). *Plasmodium malariae* infection in an asymptomatic 74-year-old Greek woman with splenomegaly. *N Engl J Med*, 338(6), 367–71.

[358] Volkman, S. K., Herman, J., Lukens, A. K., & Hartl, D. L. (2017). Genome-wide association studies of drug-resistance determinants. *Trends in Parasitology*, 33(3), 214–230.

[359] Volkman, S. K., Sabeti, P. C., DeCaprio, D., Neafsey, D. E., Schaffner, S. F., Milner Jr, D. A., Daily, J. P., Sarr, O., Ndiaye, D., Ndir, O., Mboup, S., Duraisingh, M. T., Lukens, A., Derr, A., Stange-Thomann, N., Waggoner, S., Onofrio, R., Ziaugra, L., Mauceli, E., Gnerre, S., Jaffe, D. B., Zainoun, J., Wiegand, R. C., Birren, B. W., Hartl, D. L., Galagan, J. E., Lander, E. S., & Wirth, D. F. (2006). A genome-wide map of diversity in *Plasmodium falciparum*. *Nature Genetics*, 39, 113 EP –.

[360] Volz, J. C., Yap, A., Sisquella, X., Thompson, J. K., Lim, N. T., Whitehead, L. W., Chen, L., Lampe, M., Tham, W.-H., Wilson, D., Nebl, T., Marapana, D., Triglia, T., Wong, W., Rogers, K. L., & Cowman, A. F. (2016). Essential role of the PfRh5/PfRipr/CyRPA complex during *Plasmodium falciparum* invasion of erythrocytes. *Cell Host & Microbe*, 20(1), 60 – 71.

[361] Wahlgren, M., Goel, S., & Akhouri, R. R. (2017). Variant surface antigens of *Plasmodium falciparum* and their roles in severe malaria. *Nature Reviews Microbiology*, 15, 479 EP –.

[362] Waters, A. P., Higgins, D. G., & McCutchan, T. F. (1991). *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. *Proc Natl Acad Sci U S A*, 88(8), 3140–3144.

[363] Weiss, G. E., Gilson, P. R., Taechalertpaisarn, T., Tham, W.-H., de Jong, N. W. M., Harvey, K. L., Fowkes, F. J. I., Barlow, P. N., Rayner, J. C., Wright, G. J., Cowman, A. F., & Crabb, B. S. (2015). Revealing the sequence and resulting cellular morphology of receptor-ligand interactions during *Plasmodium falciparum* invasion of erythrocytes. *PLOS Pathogens*, 11(2), 1–25.

[364] Wellems, T. E., Walker-Jonah, A., & Panton, L. J. (1991). Genetic mapping of the chloroquine-resistance locus on *Plasmodium falciparum* chromosome 7. *Proc Natl Acad Sci U S A*, 88(8), 3382–3386.

[365] Wendler, J. P., Okombo, J., Amato, R., Miotto, O., Kiara, S. M., Mwai, L., Pole, L., O'Brien, J., Manske, M., Alcock, D., Drury, E., Sanders, M., Oyola, S. O., Malangone, C., Jyothi, D., Miles, A., Rockett, K. A., MacInnis, B. L., Marsh, K., Bejon, P., Nzila, A., & Kwiatkowski, D. P. (2014). A genome wide association study of *Plasmodium falciparum* susceptibility to 22 antimalarial drugs in Kenya. *PLoS One*, 9(5), e96486–.

[366] Westenberger, S. J., McClean, C. M., Chattopadhyay, R., Dharia, N. V., Carlton, J. M., Barnwell, J. W., Collins, W. E., Hoffman, S. L., Zhou, Y., Vinetz, J. M., & Winzeler, E. A. (2010). A systems-based analysis of *Plasmodium vivax* lifecycle transcription from human to mosquito. *PLoS Negl Trop Dis*, 4(4), e653.

[367] White, N. J. (2002). The assessment of antimalarial drug efficacy. *Trends Parasitol*, 18(10), 458–64.

[368] WHO (2017). *World Malaria Report 2017*. Geneva: World Health Organisation.

[369] Wilson, C., Serrano, A., Wasley, A., Bogenschutz, M., Shankar, A., & Wirth, D. (1989). Amplification of a gene related to mammalian *mdr* genes in drug-resistant *Plasmodium falciparum*. *Science*, 244(4909), 1184–1186.

[370] Wilson, C. M., Volkman, S. K., Thaithong, S., Martin, R. K., Kyle, D. E., Milhous, W. K., & Wirth, D. F. (1993). Amplification of *Pfmdr1* associated with mefloquine and halofantrine resistance in *Plasmodium falciparum* from Thailand. *Molecular and Biochemical Parasitology*, 57(1), 151–160.

[371] Winter, G., Kawai, S., Haeggstrom, M., Kaneko, O., von Euler, A., Kawazu, S., Palm, D., Fernandez, V., & Wahlgren, M. (2005). SURFIN is a polymorphic antigen expressed on *Plasmodium falciparum* merozoites and infected erythrocytes. *J Exp Med*, 201(11), 1853–63.

[372] Witkowski, B., Duru, V., Khim, N., Ross, L. S., Saintpierre, B., Beghain, J., Chy, S., Kim, S., Ke, S., Kloeung, N., Eam, R., Khean, C., Ken, M., Loch, K., Bouillon, A., Domergue, A., Ma, L., Bouchier, C., Leang, R., Huy, R., Nuel, G., Barale, J. C., Legrand, E., Ringwald, P., Fidock, D. A., Mercereau-Puijalon, O., Ariey, F., & Menard, D. (2017). A surrogate marker of piperaquine-resistant *Plasmodium falciparum* malaria: a phenotype-genotype association study. *Lancet Infectious Diseases*, 17(2), 174–183.

[373] Wong, W., Bai, X.-C., Sleebs, B. E., Triglia, T., Brown, A., Thompson, J. K., Jackson, K. E., Hanssen, E., Marapana, D. S., Fernandez, I. S., Ralph, S. A., Cowman, A. F., Scheres, S. H. W., & Baum, J. (2017). Mefloquine targets the *Plasmodium falciparum* 80S ribosome to inhibit protein synthesis. *Nature Microbiology*, 2, 17031 EP –.

[374] Woodrow, C. J. & White, N. J. (2017). The clinical impact of artemisinin resistance in Southeast Asia and the potential for future spread. *FEMS Microbiology Reviews*, 41(1), 34–48.

[375] Wootton, J. C., Feng, X. R., Ferdig, M. T., Cooper, R. A., Mu, J. B., Baruch, D. I., Magill, A. J., & Su, X. Z. (2002). Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature*, 418(6895), 320–323.

[376] Worldwide Antimalarial Resistance Network A. L. Dose Impact Study Group (2015). The effect of dose on the antimalarial efficacy of artemether-lumefantrine: a systematic review and pooled analysis of individual patient data. *Lancet Infect Dis*, 15(6), 692–702.

[377] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER suite: protein structure and function prediction. *Nat Methods*, 12(1), 7–8.

[378] Yuan, L., Wang, Y., Parker, D. M., Gupta, B., Yang, Z., Liu, H., Fan, Q., Cao, Y., Xiao, Y., Lee, M.-c., Zhou, G., Yan, G., Baird, J. K., & Cui, L. (2015). Therapeutic responses of *Plasmodium vivax* malaria to chloroquine and primaquine treatment in Northeastern Myanmar. *Antimicrobial Agents and Chemotherapy*, 59(2), 1230–1235.

[379] Zerbino, D. R. & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829.

[380] Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, 69 Suppl 8, 108–17.

[381] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, 9, 40.

[382] Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33(7), 2302–9.

[383] Zimin, A. V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21), 2669–77.

[384] Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., Yorke, J. A., Dvořák, J., & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, 27(5), 787–792.